Panorama do tema qualidade de vida nas publicações científicas em oncologia nos últimos 10 anos: um estudo de análise de redes e processamento de linguagem natural

Overview of the quality of life theme in scientific publications in oncology in the last 10 years: a network analysis and natural language processing study

DOI:10.34119/bjhrv4n2-415

Recebimento dos originais: 04/02/2021
Aceitação para publicação: 15/03/2021

Bruno Santos Wance de Souza
Graduação em Medicina (UFRJ). MBA Executivo em Business Analytics e Big Data pela Fundação Getúlio Vargas (FGV).
Instituição atual: Hospital Sírio Libanês, Brasília
End: Condomínio Quinta Bela Vista, F12, Jardim Botânico, Brasília - DF
E-mail: bruno.swsouza@gmail.com

Diego Luis Pereira de Oliveira
MBA Executivo em Business Analytics e Big Data pela Fundação Getúlio Vargas (FGV).
Instituição de atuação atual: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)
Endereço: Setor Bancário Norte (SBN), Quadra 2, Bloco L, Lote 06, Edifício Capes
E-mail: diegolpo2000@gmail.com

Hugo Azevedo Bergamaschi
MBA Executivo em Business Analytics e Big Data pela Fundação Getúlio Vargas (FGV).
Instituição de atuação atual: Centro de Ensino Unificado de Brasília - CEUB
Endereço: 707/907 - Campus Universitário, SEPN - Asa Norte, Brasília - DF
E-mail: hugo.bergamaschi@gmail.com

André Luiz Lopes de Azevedo
MBA Executivo em Business Analytics e Big Data pela Fundação Getúlio Vargas (FGV).
Instituição de atuação atual: Centro de Ensino Unificado de Brasília - CEUB
Endereço: 707/907 - Campus Universitário, SEPN - Asa Norte, Brasília - DF
E-mail: andre.lla@outlook.com

Lucas de Jesus Matias
MBA Executivo em Business Analytics e Big Data pela Fundação Getúlio Vargas (FGV).
Instituição de atuação atual: Banco de Brasília
Endereço: rua 30 Norte Lote 01, Águas Claras, Brasília
E-mail: lucas.matias.87@gmail.com

João Tiburcio Dias de Oliveira
Doutorado em Física (UFSM) e MBA Executivo em Business Analytics e Big Data pela Fundação Getúlio Vargas (FGV)
Instituição de atuação atual: CAPES
Endereço: QI 31, lotes 2 e 4, apto 4 - 407, Guará II, Brasilia - DF
E-mail: joao.oliveir@gmail.com

Gustavo Corrêa Mirapalheta
Graduação em Engenharia Elétrica (UFRGS) e Doutorado em Administração de Empresas (FGV)
Instituição de atuação atual: EAESP/FGV
Endereço: Rua Itapeva 474 9o andar, Bela Vista, São Paulo/SP

RESUMO
Existem duas formas de abordagem no tratamento do cancer: a procura pela cura ou pela contenção da doença. Enquanto a cura procura uma forma de eliminar a doença, a contenção concentra-se em preserver a qualidade de vida do paciente. Dada a importância crescente do cancer para o setor de saude e o impacto que esta doença tem em paises de baixa e media renda, faz-se necessário entender como os tratamentos tem sido recomendados e aplicados pela comunidade médica ao longo da última década. Para tal, é feita uma análise bibliométrica da produção científica mundial, dos últimos dez anos, relacionada à área de qualidade de vida em publicações em oncologia por meio de análises de texto exploratórias (mineração de texto) e das redes de citações. Por ultimo é apresentado um método para seleção de artigos de maior influência baseado em processamento de linguagem natural.

Palavras-chaves: Câncer, Qualidade de Vida, Oncologia, Bibliometria, Processamento de linguagem natural

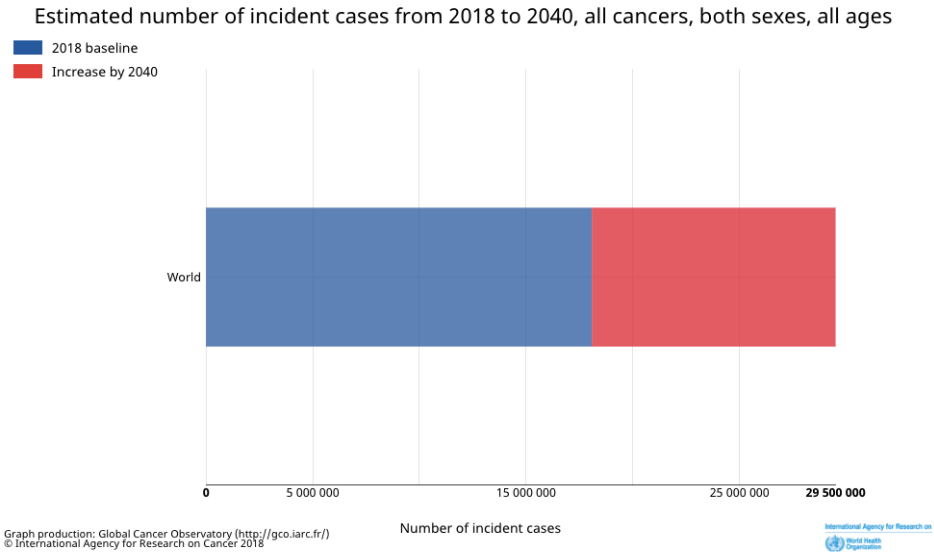
ABSTRACT
There are two ways of approaching cancer treatment: the search for a cure or the containment of the disease. While cure looks for a way to eliminate the disease, containment focuses on preserving the patient's quality of life. Given the growing importance of cancer for the health sector and the impact that this disease has in low and middle income countries, it is necessary to understand how treatments have been recommended and applied by the medical community over the last decade. To this end, a bibliometric analysis of the world scientific production in the last ten years related to the area of quality of life in oncology publications is carried out by means of exploratory text mining and citation network analysis. Finally, a method for selecting the most influential articles based on natural language processing is presented.

Keywords: Cancer, Quality of Life, Oncology, Bibliometrics, Natural Language Processing.

1 INTRODUÇÃO

O câncer é uma doença de elevada e crescente relevância mundial, afetando jovens e idosos, pobres e ricos, homens, mulheres e crianças. Em 2018 a incidência mundial de câncer foi de 18,1 milhões, enquanto projeções da Organização Mundial de Saúde (OMS) estimam que em 2040 esse número pode superar 29 milhões de casos (figura 1). O impacto econômico relacionado ao câncer é igualmente significativo e também apresenta crescimento, tendo sido estimado em aproximadamente $1,16 trilhão no ano de 2010 (1, 2).

FIGURA 1 - Estimativa do crescimento de casos de cancer entre 2018 a 2040.



FONTE: Global Cancer Observatory (http://gco.iarc.fr)

Quando diagnosticado de forma precoce o tratamento do câncer tem como objetivo a cura, no entanto em casos mais avançados o principal objetivo se torna o controle da doença. Seja qual for o estágio em que a doença se encontre, pacientes portadores de doença oncológica geralmente experimentam grande impacto negativo em qualidade de vida, com sintomas relacionados diretamente à própria doença e/ou aos tratamentos antineoplásicos necessários. Dessa forma, se torna fundamental o desenvolvimento de intervenções para o manejo adequado de sintomas e melhora da qualidade de vida em portadores de câncer (3)

Apesar de afetar a todos, estima-se que 70% das mortes por câncer aconteçam em países de baixa ou média renda, e ainda que apenas 1 em cada 5 desses países possuem dados necessários para direcionar políticas em saúde (2). Diversos estudos demonstram

que análises bibliométricas podem ajudar a analisar o estado da arte e identificar tendências, além de direcionar pesquisa e políticas relacionadas ao câncer (4).

O objetivo principal do presente estudo é apresentar o panorama dos últimos 10 anos da produção científica mundial relacionada à área de qualidade de vida em publicações em oncologia por meio de análises exploratórias e análise das redes de citações. Além disso, apresentamos um método para seleção de artigos de maior influência baseado em processamento de linguagem natural.

2 METODOLOGIA

Os metadados das publicações científicas foram baixados do *pubmed* (5) nos dias 27 e 28 de maio de 2020 usando a biblioteca especializada *pubmedR* (6). O *pubmed* é um buscador de acesso livre alimentado pela base de dados de resumos e referências de tópicos relacionados às ciências da vida e biomédicas conhecida como MEDLINE (7) e é mantido pela Biblioteca Nacional de Medicina dos Estados Unidos (8). Este foi escolhido como a fonte dos dados deste trabalho por ser uma base de dados reconhecida pela sua excelência na área médica bem como por ser de acesso aberto.

Foram baixados os metadados de artigos dos últimos 10 anos (2010-2019) relacionados à oncologia através da seguinte busca (query): *"neoplasms"[MeSH Terms] OR "neoplasms"[All Fields] OR "oncology"[All Fields] OR "oncology s"[All Fields] AND english[LA] AND Journal Article[PT] AND 2010:2019[DP]*. Devido ao significativo volume de dados resultante da busca supracitada, optamos por baixá-los de forma anualizada e posteriormente consolidá-los em uma base de dados única.

Após análise inicial da base de dados, percebemos que as referências dos artigos não estavam incorporadas aos metadados. Como essas informações seriam de primordial importância para a realização das análises propostas, utilizamos API do pubmed para buscar identificadores dos artigos que citaram aqueles presentes na base de interesse, e assim enriquecê-la com as citações.

A fim de determinar o impacto das publicações, optamos por enriquecer também a base de dados com os dados referentes aos fatores de impacto das publicações oriundos do JCR (9) e do Scimago Journal Rank (10) referentes ao ano de 2018.

A higienização dos dados foi realizada para verificação de consistência, remoção de campos desnecessários e duplicidades, além de aplicação de filtros em consonância com os objetivos do estudo. Utilizamos as informações do *MESH* (11) para selecionar apenas estudos relacionados a humanos, e também para identificar subconjunto de

publicações relacionadas à qualidade de vida, através do uso da expressão *"quality of life"*. Chamaremos as bases finais de *câncer* (939.655 publicações relacionadas à oncologia e humanos, publicadas nos últimos 10 anos) e *QoL* (20.590 publicações, subconjunto da base *oncologia* encontrados com MESH "*quality of life*") totalizando os 960.245 registros que compõem a *base completa* (*cancer = base completa - QoL*).

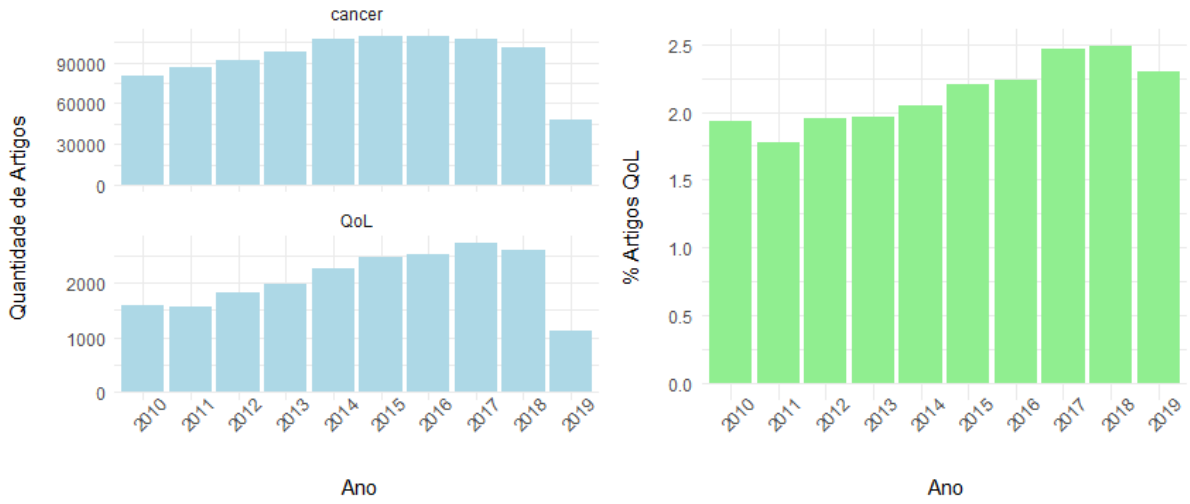
3 DISCUSSÃO E RESULTADOS

3.1 ANÁLISES DESCRITIVAS

Ao observar a base de dados *cancer*, percebemos rapidamente que a produção científica na área de oncologia é bastante significativa e apresenta tendência de crescimento ao longo da última década (figura 2, A). No ano de 2010 foram 80.320 artigos publicados, e em 2016 esse número já havia subido para 109.463 publicações no ano. Nos últimos 3 anos observamos decréscimo no número de publicações, podendo estar relacionado à variações na indexação de periódicos ao *pubmed*. Apesar de não acreditamos que tal observação esteja relacionada à redução real no número de publicações científicas em oncologia, apenas através da realização de estudo dedicado, fora do escopo do presente trabalho, poderíamos explicar melhor tal observação.

A exploração da base de dados *QoL* nos permitiu observar também uma tendência de crescimento no número de publicações relacionadas à qualidade de vida e câncer nos últimos 10 anos. Em 2010 foram publicados 1.584 artigos científicos e em 2018 esse número já supera 2.500 artigos publicados no ano (figura 2, B). Além disso, quando observamos a participação percentual do subconjunto *QoL* em relação ao conjunto total dos dados (*cancer + QoL*) percebemos também tendência de aumento (figura 2, C). Isso nos permite inferir que o tema qualidade de vida vem ganhando maior interesse e atenção dentro das publicações em oncologia.

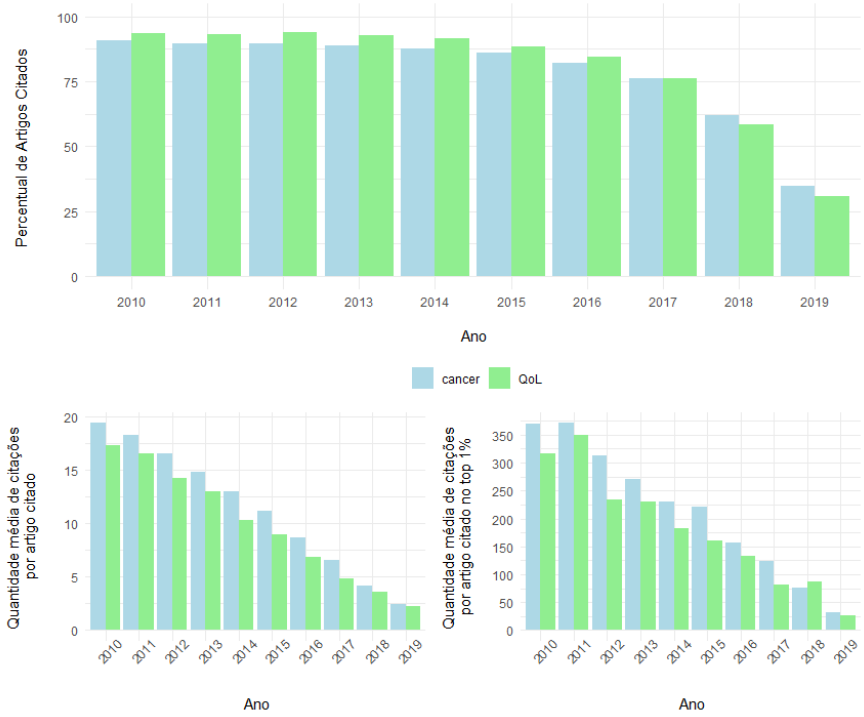
FIGURA 2 - A - Esquerda Superior) Evolução anual dos artigos publicados da base de dados *cancer*. B - Esquerda Inferior) Evolução anual dos artigos publicados na base de dados *QoL*. C - Direita) Evolução da participação percentual dos artigos *QoL* na base de dados completa (QoL + câncer).



Partindo para as análises de citações, percebemos que o percentual de artigos citados em ambas as bases *câncer* e *QoL* é bastante elevado, flutuando na faixa de 80 a 90% entre 2010 e 2016 (figura 3, A). Observamos ainda uma tendência de maior percentual de citação anual a favor dos artigos da base *QoL* quando comparados com os da base *cancer*. Nos últimos 3 anos analisados há um decrescimento esperado no percentual de citações, já que existe um tempo natural entre a publicação do artigo e o recebimento de uma citação. Nas figuras 3 B e C observamos respectivamente as quantidades médias de citações por artigo citado em ambas as bases, e análise específica para os artigos acima do percentil 99% em quantidade média de citações de cada ano (chamados de Top 1%).

Embora não exista consenso para classificar artigos como altamente citados, há abordagens que definem estes artigos como aqueles que são muito destoantes em relação a média (12) bem como abordagens baseadas nos percentis da distribuição das citações (13, 14). O uso dos percentis é devido aos artigos altamente citados serem tão extremos que influenciam na posição da média. No presente caso, percebe-se que os artigos no Top 1% chegam a ter mais de 300 citações, em média e em 2010, para ambos os conjuntos de dados conforme apresentado na figura 3 (A e B) o que seria em torno de 17 vezes a média geral do período.

FIGURA 3 - A - Superior) Percentual anual de artigos citados em cada base de dados. B - Inferior Esquerda) Quantidade média anual de citações por artigo citado em cada base de dados. C - Inferior Direita) Quantidade média anual de citações por artigo citado em cada base de dados para artigos acima do percentil 99%.



Buscamos também identificar a origem das citações, e para isso foram escolhidas 3 origens distintas: 1) base *Qol*; 2) base *cancer*; e 3) origem externa (todo artigo não pertencente à base completa: *QoL + câncer*). Nosso interesse em tal análise está em entender melhor se os temas estudados são de interesse restrito ou amplo na comunidade científica.

A origem das citações dos artigos está apresentada na figura 4. Percebemos que ambas as bases *câncer* e *QoL* recebem significativo percentual de citações de origem externa, evidenciando interdisciplinaridade e amplo interesse da comunidade científica nos temas em questão (figura 4, A e B). Observamos ainda que o percentual de citações externas vem aumentando nos últimos anos possivelmente devido às citações provenientes de áreas cujo ciclo publicação-citação seja mais curto do que a oncologia ou devido a um aumento no interesse por publicações desta área. O quanto cada uma dessas possibilidades suporta o comportamento observado está além do escopo do presente estudo e não dispomos de evidência suficiente para suportar essas hipóteses. O percentual de citações externas dos artigos altamente citados (figuras 4, C e D) é ainda maior em ambas as bases, o que faz sentido se considerarmos que uma das forma de um artigo

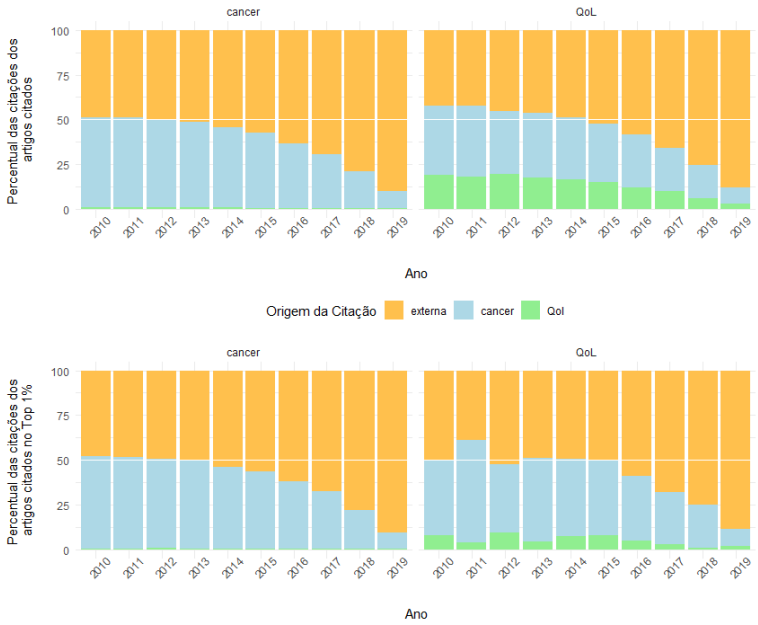
atingir elevado número de citações é cruzar as fronteiras de uma área específica do conhecimento científico.

TABELA 1 - Esquerda) Top 10 publicações em *cancer*. Direita) Top 10 publicações em *QoL*.

Publicação	**N**	**%**
PLoS ONE	21.253	2.26 %
Oncotarget	12.724	1.35 %
BMC Cancer	7.633	0.81 %
Asian Pac. J. Cancer Prev.	7.366	0.78 %
Anticancer Res.	7.339	0.78 %
Sci Rep	7.075	0.75 %
Oncol. Rep.	6.190	0.66 %
Clin. Cancer Res.	6.154	0.65 %
Tumour Biol.	5.694	0.61 %
Medicine (Baltimore)	5.694	0.61 %
	87.122	**9.27 %**

Publicação	**N**	**%**
Support Care Cancer	1.023	4.97 %
Psychooncology	701	3.4 %
Qual Life Res	414	2.01 %
BMC Cancer	385	1.87 %
Cancer	285	1.38 %
J. Clin. Oncol.	284	1.38 %
Eur J Cancer Care (Engl)	252	1.22 %
Eur J Oncol Nurs	247	1.2 %
Cancer Nurs	240	1.17 %
Health Qual Life Outcomes	235	1.14 %
	4.066	**19.75 %**

FIGURA 4 - A - Superior) Origem das citações para cada uma das bases de dados em geral e B - Inferior) para artigos situados no percentil acima de 99%.



Identificamos os 10 veículos responsáveis pelo maior número de publicações científicas em ambas as bases (Top 10), e percebemos que se tratam de veículos diferentes. Apenas o periódico *BMC Cancer* está presente no Top 10 de ambas as bases. Observamos ainda que o Top 10 da base *QoL* é responsável por 19,75% de todas as publicações da área, enquanto na base *cancer* esta participação é de apenas 9,27%. Gostaríamos de ressaltar também os títulos dos veículos presentes no Top 10 da base

QoL: observamos veículos direcionados para psico-oncologia e enfermagem, sendo essa mais uma constatação da interdisciplinaridade do tema qualidade de vida em oncologia.

3.2 PROCESSAMENTO DE LINGUAGEM NATURAL

O uso de Processamento de Linguagem Natural (PNL) na área da medicina está em franco desenvolvimento. Recente revisão sistemática da literatura evidenciou cerca de 100 publicações anuais nos últimos 5 anos, sendo câncer o principal assunto estudado, correspondendo a aproximadamente 25% de todas essas pesquisas (15).

Devido ao já demonstrado aumento do número de publicações científicas em oncologia, a tarefa de se manter atualizado vem se tornando cada vez mais desafiadora para os pesquisadores e profissionais da área. Técnicas de gerenciamento de artigos científicos utilizando PNL podem contribuir para a prática diária do pesquisador, ajudando o mesmo a entender o estado da arte e identificar tendências, além de priorizar, por exemplo, os estudos mais influentes da área de pesquisa desejada através de ranqueamento de artigos.

No presente estudo, aplicamos técnicas de PNL sobre a base de dados *QoL* para: 1) identificar os termos mais frequentes presentes nas publicações; 2) identificar identidades geográficas, culturais e organizacionais; 3) criar ranking de publicações científicas mais influentes em subáreas de pesquisa; e 4) realizar o agrupamento dos artigos que guardem mais similaridades entre si.

3.3 IDENTIFICAÇÃO DE TERMOS MAIS FREQUENTES.

A identificação dos termos mais frequentes presentes nas publicações, foi feita após fase inicial de processamento dos dados. Nessa fase, foram realizadas dentre outros procedimentos tratamento dos campos faltantes, lematização e remoção das *stop words* (16), palavras que não contribuem para análises de PNL. Por não existir base universal de *stop words* escolhemos utilizar o modelo SpaCy (17), pacote específico para processamento de textos biomédicos. A figura 5 mostra o mapa de palavras criado com os termos mais frequentes encontrados.

FIGURA 5 - Mapa de palavras contendo os termos mais frequentes da base de dados *QoL*

Analisando a nuvem de palavras(figura 5), podemos perceber que estão entre os termos mais frequentes, os de "Pacientes com Câncer" (*cancer patient*), "Câncer de Mama" (*breast cancer*), "Câncer de Próstata*"* (*prostate cancer*) e "Sobreviventes de Câncer" (*cancer survivor*). Nesse contexto, vale destacar que esses termos foram fundamentais para que se fosse feito uma pesquisa específica em nossa base de dados, a fim de descobrirmos quais são as publicações relacionadas a esses termos mais bem ranqueadas tanto no JCR quanto no SJR.

3.4 IDENTIFICAÇÃO DE IDENTIDADES CULTURAIS.

Por se tratar de base de dados global, é esperado que as publicações científicas apresentem diferentes identidades geográficas, culturais, e organizacionais relacionadas ao contexto no qual estejam inseridas. Nesse sentido, a figura 6 apresenta um exemplo relacionado ao reconhecimento de entidades nomeadas (*NER*), isto é, aos termos gerais que se repetem constantemente nos textos, porém, apresentando significados semânticos relevantes.

FIGURA 6 - Exemplo de reconhecimento de entidades nomeadas no abstract de um artigo da base de dados *QoL*.

Cabe observar que para a análise do NER usamos o, já mencionado, pacote *SpaCy* para determinar as principais identidades em 3 distintos domínios: 1) Nacionalidades, grupos religiosos ou políticos (NORP); 2) empresas, agências, instituições e outras organizações (ORG); e 3) países, cidades, estados (GPE) (18). Foram selecionadas apenas as entidades com mais de 50 aparições e as figuras 7 e 8 mostram os 3 domínios em questão.

FIGURA 7 - Esquerda) Identidades identificadas no domínio Cultural - Nacionalidades, grupos religiosos ou políticos (NORP). Direita) Identidades identificadas no domínio Geográfico - países, cidades, estados (GPE).

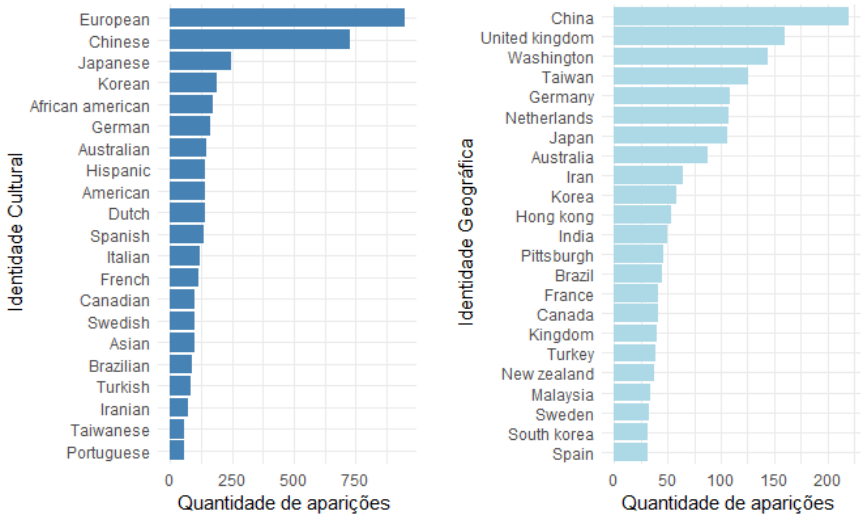
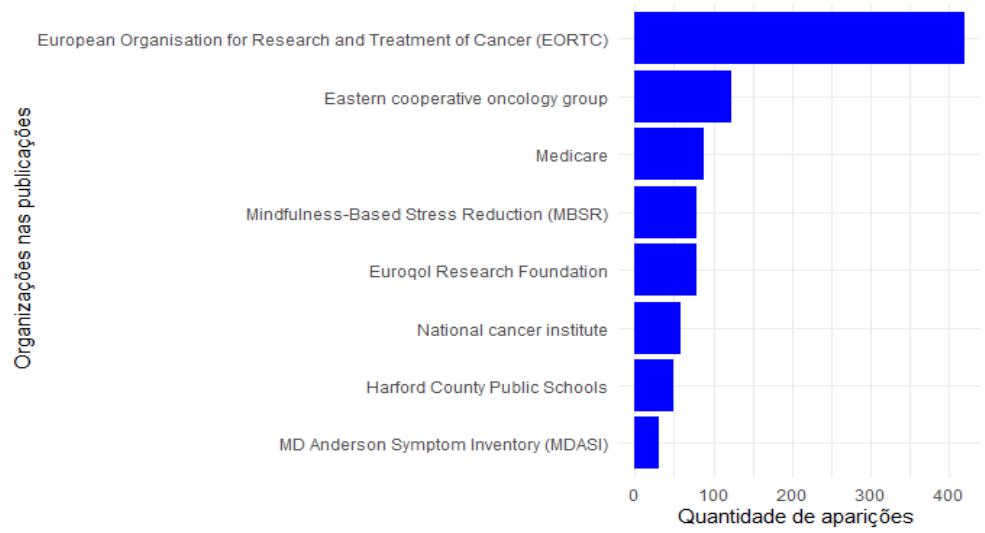


FIGURA 8 - Identidades identificadas no domínio das Organizações Públicas - empresas, agências, instituições e outras organizações (ORG).



3.5 RANKING DE PUBLICAÇÕES CIENTÍFICAS MAIS INFLUENTES EM SUBÁREAS DE PESQUISA.

Para desenvolvimento do sistema de rankeamento de publicações mais influentes em determinada subárea de pesquisa optamos por avaliar artigos relacionados aos sobreviventes do câncer de mama (*breast cancer survivor*) e sobreviventes do câncer de próstata (*prostate cancer survivor*). A escolha das subáreas em questão se deu por serem as entidades oncológicas de maior incidência mundial em mulheres e homens, e com grande impacto na qualidade de vida de pacientes tratados e sobreviventes, conforme verificado no mapa de palavras.

Para identificar artigos relacionados à subárea específica utilizamos a grandeza *TF* e *IDF*, já que a mesma representa a importância de um termo para um documento inserido em uma coleção ou corpus. Além disso, cabe ressaltar que foi aplicado a *Similaridade do cosseno nos vetores* (19), a julgar pelo fato de que para cada artigo, derivamos um vetor e vetores lidam apenas com números (20). Nesse sentido, vale ressaltar que no presente trabalho estamos lidando apenas com documentos de textos, por isso, essa foi a razão pela qual usamos *TF* e *IDF*, ou seja, para converter nos artigos em números, a fim de que eles possam ser representados por um determinado vetor. Afinal, o conjunto de artigos em nossa coleção pode ser compreendido como um conjunto de vetores em um espaço vetorial; consequentemente, cada termo terá seu próprio eixo. Desse modo, cabe dizer que essa métrica de similaridade depende da visualização das preferências do usuário como pontos no espaço. Em síntese, cabe frisar que através do

TF-IDF com a *similaridade de cosseno*, nós conseguimos detectar a similaridade entre nossos artigos em relação a toda nossa base de dados.

Os anexos 1 e 2 trazem a lista dos 30 artigos mais influentes relacionados respectivamente com pesquisas de qualidade de vida em pacientes sobreviventes de câncer de mama e de próstata. Para cada artigo foi atribuída sua avaliação de impacto no JCR e SJR, além de similaridade com os termos de busca utilizados. Como resultado, nota-se que nessa base de dados sobre oncologia e, especificamente, sobre qualidade de vida dos indivíduos, a abordagem referente aos sobreviventes do câncer de mama, teve um maior impacto - de aproximadamente 61% - no artigo *Relationships between cause of cancer and breast cancer-related factors in breast cancer survivors*, sem classificação tanto em JCR quanto em SJR. Não obstante, podemos observar ainda que o artigo com maior fator de impacto, tanto do JCR (4,011) quanto do SJR(1,414) é o *"Being slightly overweight is associated with a better quality of life in breast cancer survivors"*, o qual obteve aproximadamente 54% de impacto relacionado ao tema pesquisado.

3.6 AGRUPAMENTO DE ARTIGOS SIMILARES

Nessa etapa, nós buscamos uma maior eficiência no gerenciamento da nossa base de dados, tendo em vista que agrupamos os artigos que mais guardam similaridade entre si, a julgar pelo fato de que isso facilitaria nossa pesquisa em buscar um grupo específico de determinado conteúdo abordado. Sendo assim, nós executamos 3 (três) etapas relevantes, quais sejam:

- Redução de dimensionalidade utilizando o *PCA (Principal Component Analysis)*;
- Definição do número de clusters:
    - Primeiro, pelo *Elbow Method e a métrica WCSS*;
    - Segundo, refinando a escolha do grupo através da *análise de silhueta*.
- Agrupamento através do *Fast k-means*.
-

3.6.1 Redução de dimensionalidade utilizando o PCA (Principal Component Analysis)

Ao levar em conta de que lidamos com uma base de dados de tamanho extenso, foi necessário realizar uma diminuição de dimensionalidade para lidar com os atributos totais. Com isso, foi aplicado a referida redução através do PCA *(Principal Component Analysis)*, tendo em vista ser um dos principais algoritmos de aprendizagem de máquina não supervisionada capaz de lidar com uma base de dados altamente assimétrica (dados

distribuídos irregularmente) como é o caso em análise. Além disso, cabe observar que o referido algoritmo busca identificar a correlação entre as variáveis, e, caso haja uma forte correlação, então, torna-se possível reduzir a dimensionalidade da base.

Em linha gerais, o formalismo algébrico que representa esse conceito, segue a seguinte intuição:

- Das "*m"* variáveis independentes, PCA extrai "*p<=m"* novas variáveis independentes que melhor explique a variação na nossa base de dados, sem considerar a variável dependente, logo, ele não considera a classe. Assim, frisa-se que uma vantagem na utilização desse algoritmo, deve-se ao fato de que podemos escolher o número de "*p"*.

Portanto, o objetivo de fazermos essa implementação do PCA, justifica-se pelo fato de que nossa base apresentou uma enorme variedade de característica para cada publicação, de modo que esses dados precisavam ser representados em um plano cartesiano, ou seja, com o valor de X e com o valor de Y; por consequência disso, ao invés de representarmos cada publicação pelas suas 25 (vinte e cinco) características, nós decidimos por representá-las por apenas 2 (duas) características (anexo 3).

3.6.2 Definição do número de clusters: Elbow Method e a métrica WCSS.

Dada a variedade de características para cada artigo, consideramos que não haveria uma garantia para encontrarmos o melhor conjunto de clusters. Sendo assim, optou-se pelo uso do *"Elbow Method"* com a métrica *"Within-cluster sum of squares" (WCSS)*, por entendermos que o método poderia se adequar a nossa base de dados e apontar um número de cluster eficiente, ou seja, o próprio modelo iria testar a quantidade de cluster que melhor se adequaria a nossa base de dados (21). Contudo, a métrica do WCSS estabelece que quanto maior o seu valor, pior será a escolha dos clusters, por isso, escolhe-se aquele que apresenta uma amenização da queda acentuada no gráfico. Por consequência disso, a referida situação gerou uma incerteza na escolha entre os números 10 (dez) e 17 (dezessete). Essa observação pode ser verificado no Anexo 4 deste artigo.

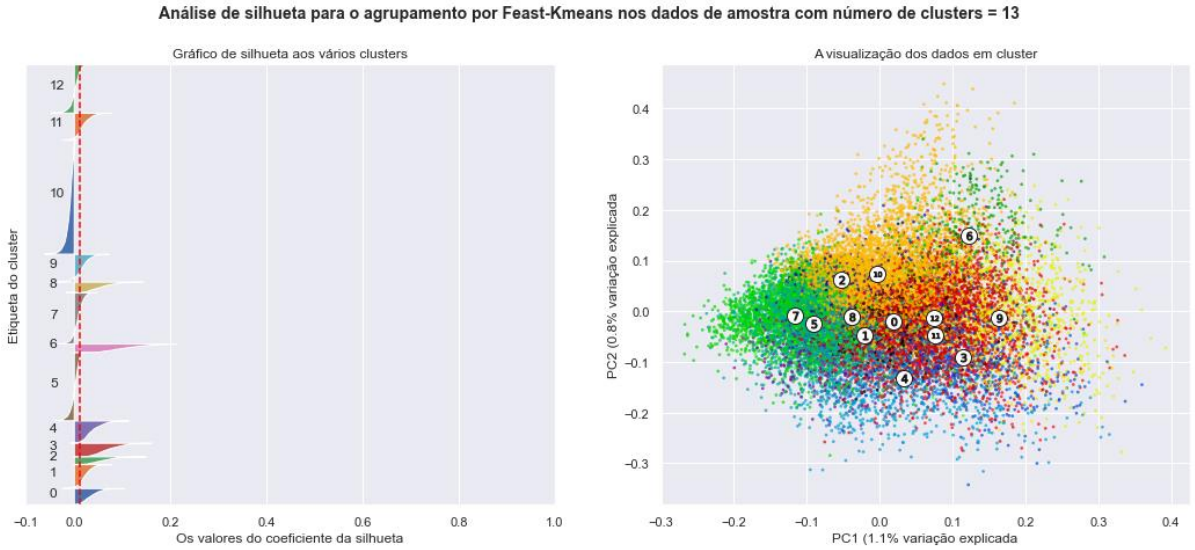
3.6.3 Definição do número de clusters: Análise de Silhueta.

Com a finalidade de termos mais solidez quanto à escolha de número de cluster, cabe destacar que através da informação trazida pela métrica *WCSS*, a saber, de que a melhor escolha para o número ideal de grupos encontra-se entre 10 e 17, foi aplicado a técnica de *análise de silhueta (22)*. Com isso, foi possível refinar ainda mais a escolha

quanto ao número de clusters, ou ainda, tornando esse processo de escolha mais eficiente. Ao final do processo, foi possível agrupar os artigos de nossa base de dados em grupos com artigos similares (Anexo 5).

A escolha do número de cluster através da análise de silhueta segue algumas diretrizes específicas, uma vez que leva em consideração tanto a medida de distância de um ponto para todos os pontos do mesmo grupo de artigos (*coesão*), quanto à medida de distância de um ponto com os pontos dos outros clusters (*separação*). Além disso, torna-se necessário que o coeficiente de aceitação ou não do cluster esteja no intervalo dos valores -1 e 1, sendo que o valor -1 indica que o cluster está ruim, por outro lado, o valor 1 indica que os clusters estão bem distantes, logo, tendem a ser escolhas mais eficiente (23).

FIGURA 9 - Análise de Silhueta.



A figura 9 mostra que a escolhemos o número de 13 (treze) clusters por entendermos que sua pontuação média superior aos clusters com 9, 10, 11, 12 e 14, por possuir de espessura semelhantes e, portanto, ter tamanhos semelhantes, como também, uma maior distância entre os círculos numéricos apresentados ao lado direito do gráfico em relação aos clusters com 15, 16 e 17 (Anexo 5).

3.6.4 Agrupamento através do Fast k-means

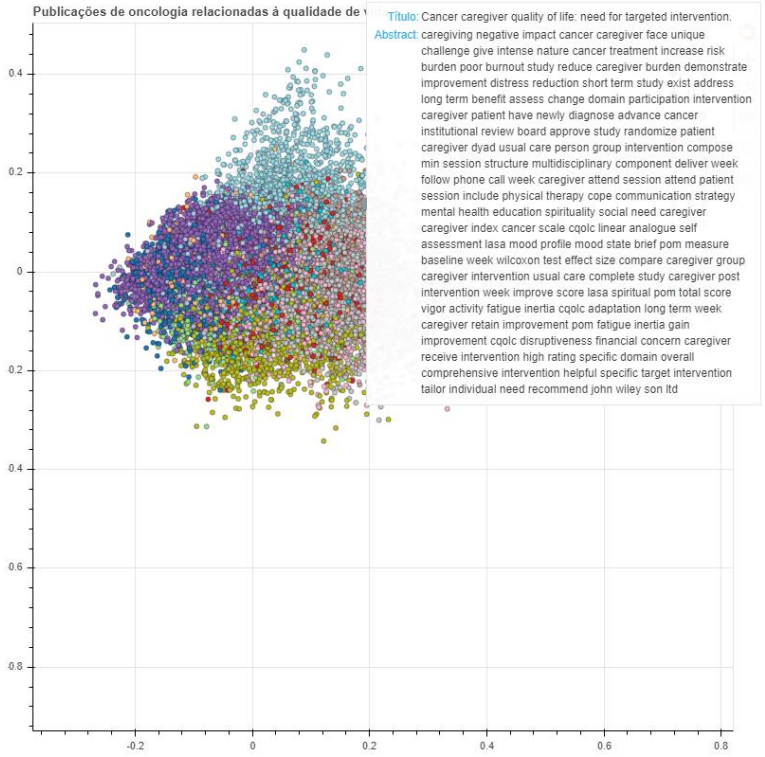
Após concluirmos tanto pelo *Elbow Method* quanto pela *Análise de Silhueta* que o número de clusters mais eficiente para agrupar os artigos similares de nossa base de dados seria 13 (treze), nós efetivamente realizamos o agrupamento dessas publicações

com a aplicação do algoritmo *Fast K-means*. A escolha desse a algoritmo, justifica-se pelo fato de que ele tem como vantagem evitar que seja escolhido pontos iniciais que geram cluster ruins, pois ele seleciona os centróides iniciais que estão longes uns dos outros; além disso, também há evidências de que para um conjunto de grande número de dados, as experiências técnicas mostram que o método de agrupamento em dois estágios alcança melhor velocidade, conforme defendido por *Kecman* e *Strack* (2011) (24). Sendo assim, decidimos fazer uso desse algoritmo, principalmente por apresentar uma velocidade de processamento mais rápida que o *K-means* tradicional.

FIGURA 10 - Agrupamento das publicações semelhantes de Oncologia direcionadas à Qualidade de Vida

Conforme gráfico apresentado, temos que as cores apresentam a semelhança entre os artigos da nossa referida base de dados, ou seja, a forma como eles estão devidamente agrupados. Ademais, ao verificar cada círculo, nota-se que cada publicação tem seu título e abstract para nos situarmos em relação ao grupo ao qual a mesma pertence.

FIGURA 11 - Agrupamento das publicações semelhantes de Oncologia direcionadas à Qualidade de Vida - Título e Abstract.



3.7 ANÁLISE DAS REDES DE CITAÇÕES

As estatísticas de citações podem ser entendidas como uma medida bastante geral do nível de contribuição que um indivíduo fez para a prática da ciência (25). Assim sendo, a reputação acadêmica de um indivíduo aumenta na medida em que seus artigos são citados. Ademais, também pode ser assumido que, se um pesquisador é citado por um de seus pares famosos, sua reputação também é aumentada. Portanto, ao pode-se usar as inter-relações contidas em uma rede de citações entre artigos como uma forma para se avaliar o prestígio/importância de pesquisadores e publicações (26).

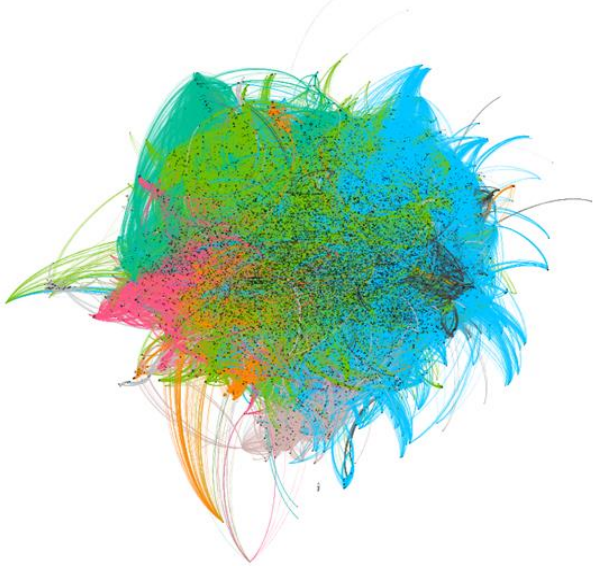
A relação entre autores ou entre publicações não é uma relação direta, mas mediada pelos artigos, ou seja, é por meio dos artigos que são feitas as citações entre os pesquisadores e entre os veículos. Além disso, essa rede de citações é uma rede direcionada com as arestas apontando do artigo citante (e portanto seus autores/publicações) para os artigos citados.

Ao tratarmos os dados como uma rede abre-se a possibilidade do uso de medidas específicas de rede. Assim, a quantidade ligações de nó (autor/publicação) medida pela *centralidade de grau* é o equivalente a medida da quantidade de citações que um indivíduo recebeu. Ademais existem medidas que levam em consideração tanto a quantidade de ligações do nó analisado quando dos nós ligantes (ou vizinhos). Uma

medida deste tipo é o *pagerank* (27) e seria o equivalente a medida de reputação em que são avaliadas tanto as citações recebidas pelo indivíduo quanto a importância relativa de quem cita.

A figura 12 apresenta a rede de citações da base de dados *QoL* com os *nós* (ou *vértices*) ocupados pelos autores e as *arestas* indicando quando há uma ligação entre os autores (por meio da citação entre artigos). Percebe-se que esta rede é bastante intrincada sendo difícil identificar visualmente a emergência de algum padrão dada a quantidade de artigos desta base de dados. Além disso, conforme mencionado e evidenciado na figura 3B, há uma disparidade muito grande na quantidade de citações entre os diferentes artigos. Estas diferenças se refletem no surgimento de comunidades (*modularidade*) definidas de acordo com a quantidade de ligações entre os diferentes autores criadas pelas citações dos artigos. No presente caso, estas comunidades estão apresentadas pelas diferentes cores da figura 12.

FIGURA 12 - Rede de citações entre autores sem tratamento.



Para a montagem destas redes, optamos por usar apenas a base de dados QoL e as citações efetuadas apenas por artigos dentro desta por considerar que esta base possui uma homogeneidade de temáticas e formas de comunicação/citação. Esta consideração foi baseada na própria especialização do tema e acabou reforçada pela constatação de que grande parte dos artigos dessa temática concentram-se em poucas publicações científicas (Tabela 1). Assim, as citações recebidas por um artigo e que foram feitas por artigos externos a base de dados *QoL* não estão sendo consideradas.

As medidas de importância relativa que usaremos são baseadas em 2 abordagens diferentes: uma usando a centralidade de grau com cada citação apresentando peso 1 e outra usando pagerank mas com cada citação apresentando como peso o fator de impacto JCR da publicação do artigo. Para ambos os casos não houve procedimento de ponderação pelo quantitativo de autores do artigo citado, ou seja, cada autor recebeu a mesma citação. Montadas as redes foram feitos os gráficos para autores e publicações e selecionados para apresentação apenas aqueles que apresentaram os 20 maiores valores da medida aplicada.

Nas figuras 13 e 14 estão apresentadas as redes obtidas para autores e publicações, respectivamente, usando a rede com peso 1 nas arestas e medindo a centralidade de grau de cada nó. As cores dos nós e arestas nas figuras 10 e 11 representam a modularidade dos nós, ou seja a formação de comunidades. Percebe-se que o autor com o maior quantitativo de citações recebidas foi o prof David Cella e a publicação mais citada foi o J. Clin. Oncol.. Os resultados dessas medidas de centralidade para os 10 autores e publicações mais bem colocados estão apresentados na Tabela 2.

FIGURA 13 - Rede de citações entre autores contendo os 20 autores com maior centralidade de grau sendo todos os artigos considerados com o mesmo peso.

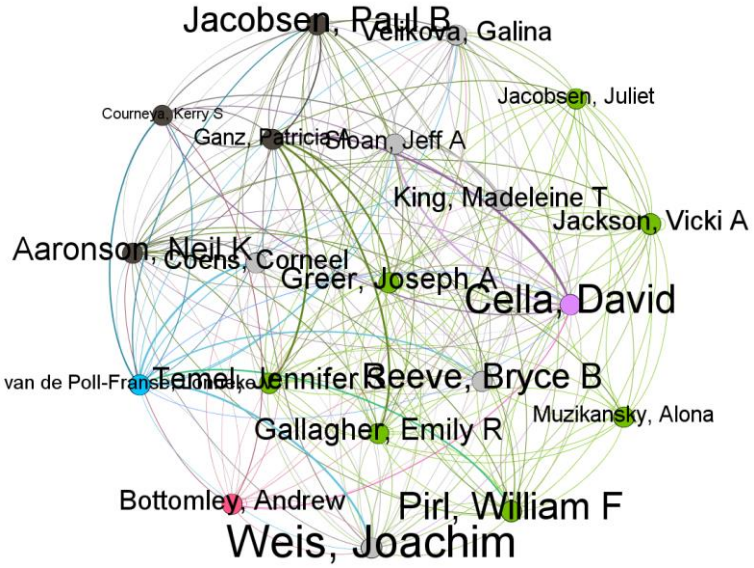


FIGURA 14 - Rede de citações entre publicações contendo as 20 publicações com maior centralidade de grau sendo todos os artigos considerados com o mesmo peso.

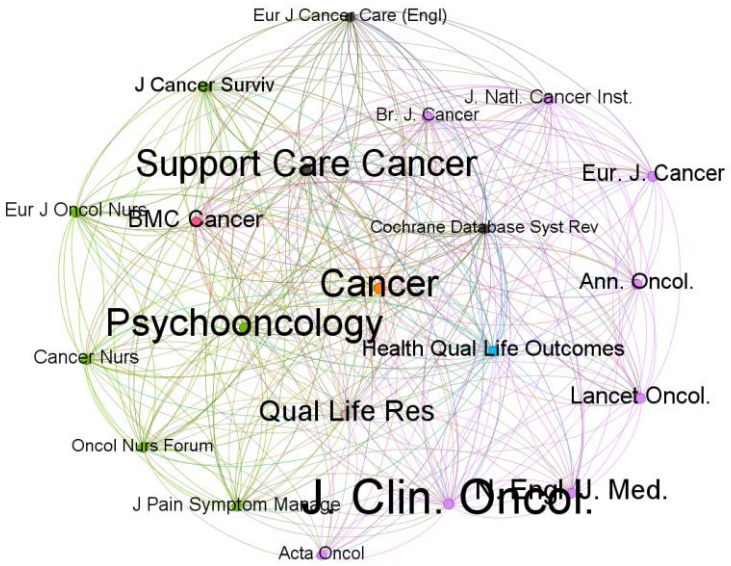


TABELA 2 - Esquerda) Os 10 autores com maior centralidade de grau apresentados na figura 13. Direita) As 10 publicações com maior centralidade de grau apresentadas na figura 14.

Pos	Autor	Centralidade de Grau	Publicação	Centralidade de Grau
1	Cella, David	2.443	J. Clin. Oncol.	272
2	Aaronson, Neil K	2.035	Psychooncology	214
3	Bottomley, Andrew	1.896	Cancer	209
4	Reeve, Bryce B	1.834	Support Care Cancer	206
5	Pirl, William F	1.748	Qual Life Res	160
6	Ganz, Patricia A	1.663	N. Engl. J. Med.	151
7	Velikova, Galina	1.580	BMC Cancer	130
8	Temel, Jennifer S	1.521	Lancet Oncol.	125
9	Greer, Joseph A	1.492	Eur. J. Cancer	121
10	Gallagher, Emily R	1.442	Ann. Oncol.	118

Na segunda abordagem para medida da importância de autores e publicações, utilizamos o fator de impacto do JCR como peso para cada uma das citações e o algoritmo de *pagerank*. Como já mencionado o *pagerank* também leva em consideração no cálculo a reputação de quem cita. No entanto, ao utilizarmos uma rede de citações reduzida

acabamos por remover as citações que alguns autores receberiam de artigos externos a área de *QoL*. Não seria exagerado inferir que os autores mais renomados seriam aqueles cujas citações são mais impactadas por esta restrição. Dessa forma, como os valores dos fatores de impacto são medidos em relação ao conjunto de artigos publicados em uma determinada publicação e em um determinado período e não apenas a um determinado artigo, ao usarmos o JCR como peso, de certa forma, estaríamos emprestando o prestígio das revistas aos autores. Assim, o uso do pagerank associado com o fator de impacto seria uma forma de magnificar a importância de algumas citações em detrimento de outras, favorecendo, em especial, aquelas provenientes das revistas mais prestigiadas. As figuras 15 e 16 apresentam as redes com os top 20 autores e publicações, respectivamente, obtidos por meio deste procedimento sendo as cores dos nós e arestas relacionada a modularidade. Um resumo contendo os 10 principais autores e publicações medidos desta forma está apresentado na tabela 3.

FIGURA 15 - Rede de citações entre autores contendo os 20 autores com maior pagerank sendo o fator de Impacto do JCR considerado como o peso da citação, que é apresentado na espessura da aresta.

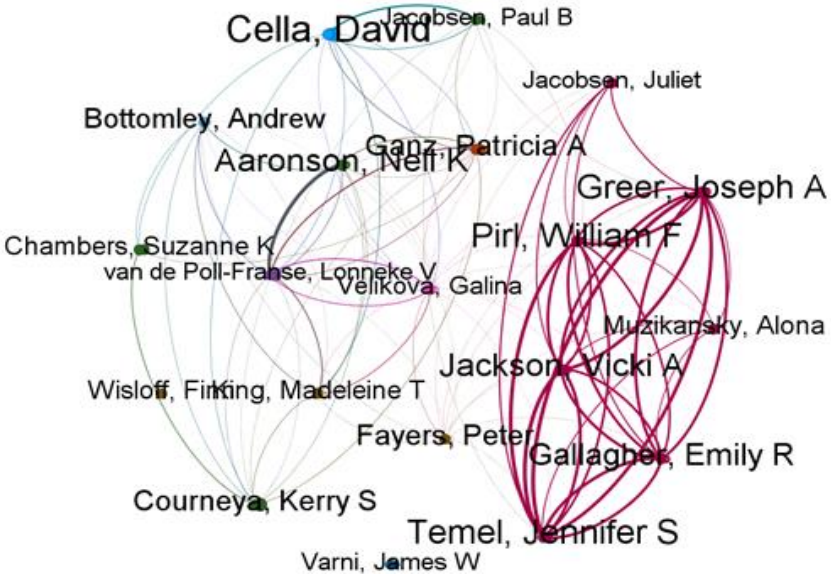


FIGURA 16 - Rede de citações entre publicações contendo as 20 publicações com maior pagerank sendo o fator de Impacto do JCR considerado como o peso da citação, que é apresentado na espessura da aresta.

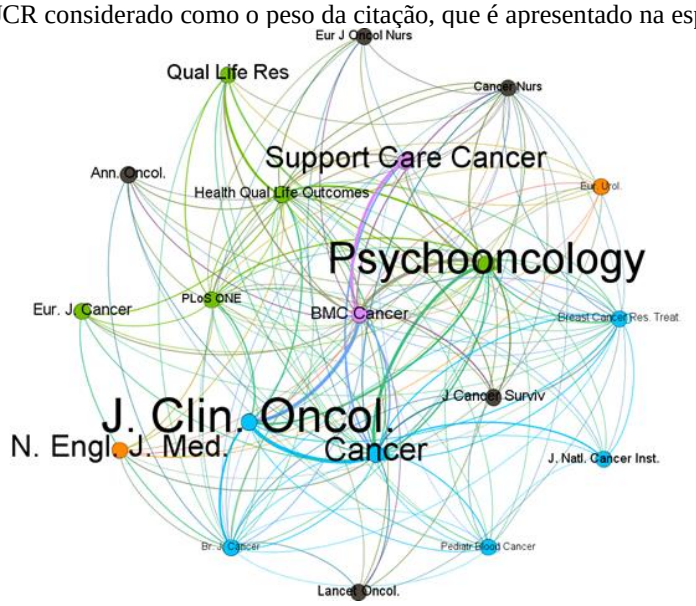
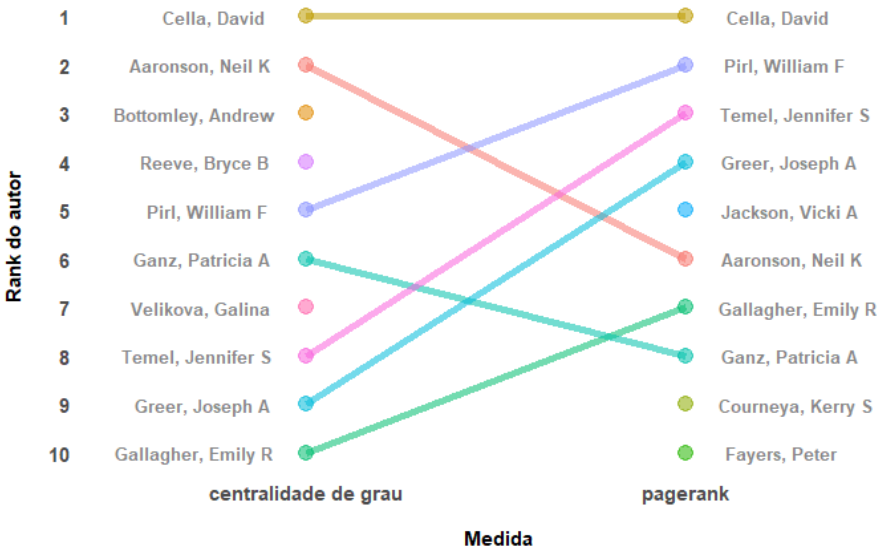


TABELA 3 - Esquerda) Os 10 autores com maior pagerank apresentados na figura 12. Direita) As 10 publicações com maior pagerank apresentadas na figura 13.

Pos	Autor	PageRank	Publicação	PageRank
1	Cella, David	0,000389	J. Clin. Oncol.	0,011357
2	Pirl, William F	0,000346	Psychooncology	0,010283
3	Temel, Jennifer S	0,000345	Cancer	0,007420
4	Greer, Joseph A	0,000338	N. Engl. J. Med.	0,007081
5	Jackson, Vicki A	0,000331	Support Care Cancer	0,006931
6	Aaronson, Neil K	0,000327	Qual Life Res	0,004365
7	Gallagher, Emily R	0,000310	BMC Cancer	0,003475
8	Ganz, Patricia A	0,000292	Eur. J. Cancer	0,003192
9	Courneya, Kerry S	0,000283	Health Qual Life Outcomes	0,003057
10	Fayers, Peter	0,000280	J Cancer Surviv	0,003043

Numa comparação entre os dados da tabela 2 com aqueles da tabela 3 percebe-se que há muitos autores e publicações que são identificados como Top 10 em ambas abordagens. Além disso, quando comparadas as publicações das tabelas 1, 2 e 3 percebe-se grande sobreposição sugerindo a existência de publicações preferenciais dentro da comunidade de *QoL*. Um resumo das posições (ou *ranks*) contidas nessas tabelas bem como suas variações de acordo com a abordagem está apresentada nas figuras 17 e 20.

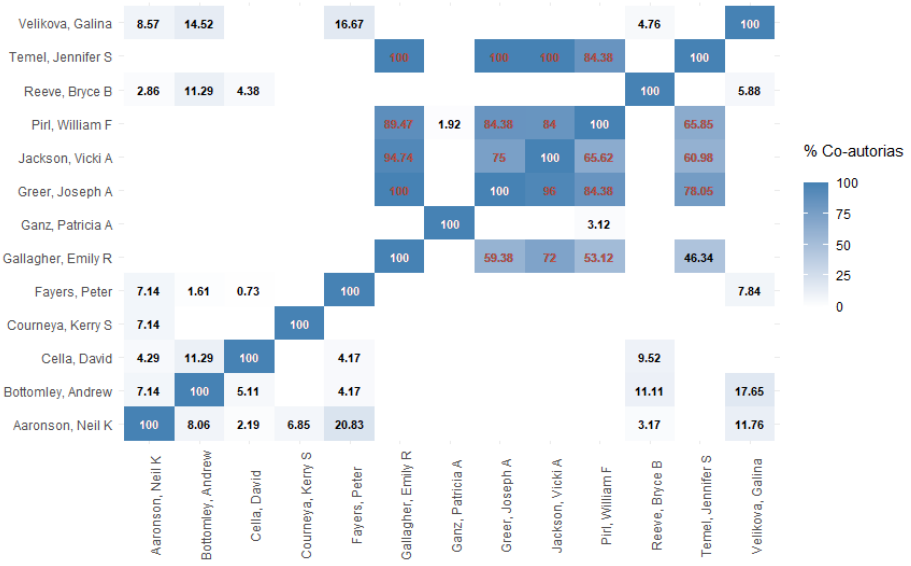
FIGURA 17 - Comparação entre as posições dos autores apresentados nas tabelas 2 e 3.



Na variação no rank dos autores dos artigos de *QoL*, figura 17, chama a atenção de que o Prof. David Cella fica na primeira posição em ambas as abordagens. Para os demais autores há variações nas posições sendo que alguns autores bem colocados na análise por quantidade de citações não aparecem entre os 10 primeiros na análise pelo pagerank. Independentemente da abordagem utilizada para identificar os principais autores, há conjunto de problemas ou dificuldades relacionados a correta contabilização da contribuição de cada autor. De uma maneira muito simplificada poderia-se apontar as auto-citações, a presença de homônimos (pessoas diferentes com mesmo identificador), sinônimos (mesma pessoa mas usando diferentes identificadores) e a dificuldade em atribuir créditos para artigos com múltiplos autores (28). Para este último caso, diversas são formas de tratamento em relação a atribuição de pesos tais como pesos iguais para todos os autores (a que escolhemos), peso maior para o primeiro autor, peso maior para o último autor, um peso em função da posição do autor na lista de autores, etc.. Desta forma, como investigação adicional, resolvemos averiguar as relações de co-autorias entre os

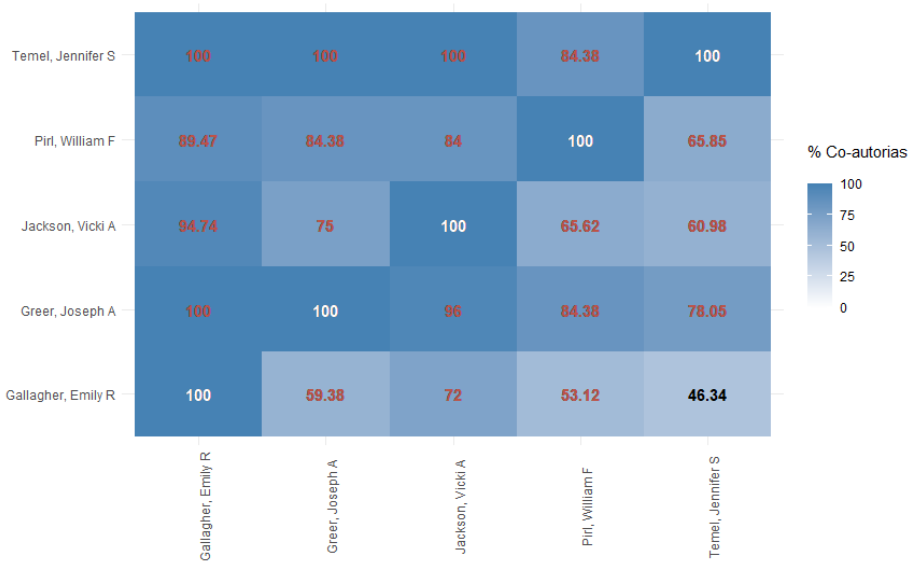
autores identificados entre os Top 10 tanto na abordagem de quantidade de citações quanto de pagerank. A figura 18 apresenta esta relação para os 13 autores selecionados.

FIGURA 18 - Percentual de artigos em que os autores apresentados nas tabelas 2 e 3 apresentaram co-autorias.



O cruzamento entre linhas e colunas na figura 18 indica o percentual de artigos de cada autor em que estes foram co-autores. Assim, esta matriz não é simétrica em relação a diagonal. Como exemplo, dos artigos de *Aaronson, Neil K.* (1º no eixo x), 7.14% foram escritos com *Bottomley, Andrew* (penúltimo no eixo y) e dos artigos de *Bottomley, Andrew*, 8.06% foram escritos com *Aaronson, Neil K.* As intersecções entre autores em branco indicam que não há artigos em que sejam co-autores e as intersecções com *labels* em vermelho indicam quando houve co-autoria em mais de 50% dos artigos. Percebe-se, claramente a existência de um *cluster* entre autores que são co-autores em um grande percentual de seus artigos sendo este um indicativo de que estes autores formam um grupo de pesquisa. Este *cluster* está apresentada de maneira isolada na figura 19. Para os demais autores, há um pequeno percentual de artigos em co-autoria de modo que pode-se concluir que não pertencem habitualmente ao mesmo grupo de pesquisa.

FIGURA 19 - Percentual de artigos em que autores selecionados apresentaram co-autorias.

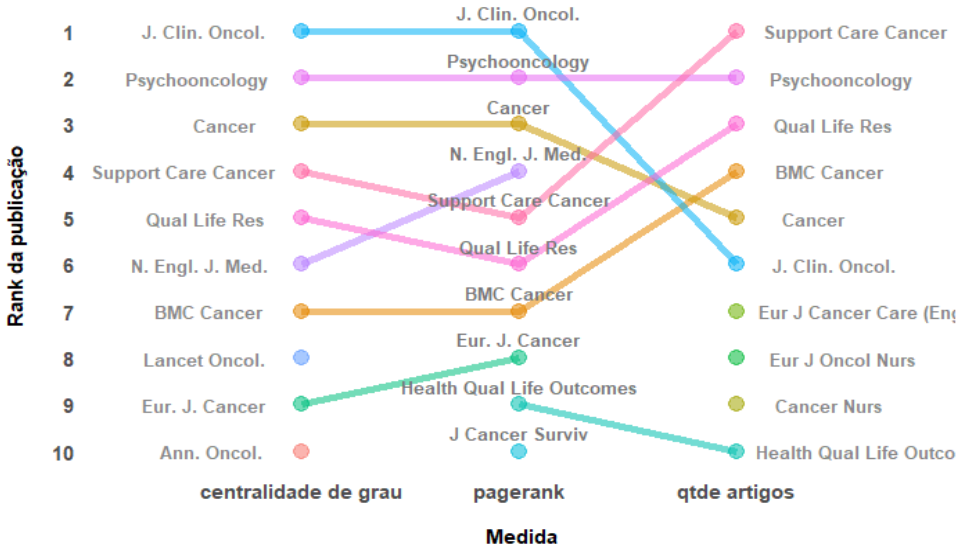


A variação no rank das publicações apresentada na figura 20 mostra esse rank medido tanto do ponto de vista de publicação quanto de citação. Chama a atenção o New England Journal of Medicine e o European Journal of Cancer terem artigos importantes do ponto de vista de citação mas não serem igualmente relevantes do ponto de vista de quantidade de artigos publicados. Isso pode ser devido a dificuldade de publicação em ambos os veículos ou a estes serem de temática mais abrangente do que aquela relacionada a *QoL*. As figuras 14 e 16 apresentam com cores diferentes para nós e vértices. Essas cores são o resultado de uma medida de modularidade, uma forma de clusterização, e indicam que há uma propensão maior a algumas ligações. Assim, a cor verde parece sugerir que os periódicos com uma temática mais voltada para o câncer e sua relação com outras áreas tendem a se citar em conjunto. Como representantes deste grupo se destacam as publicações: Psycooncology, Cancer Nursing (*Cancer Nurs*), Oncology Nursing Forum (*Oncol Nurs Forum*), European Journal of Oncology Nursing: The Official Journal of European Oncology Nursing Society (*Eur J Oncol Nurs*), Journal ff Cancer Survivorship : Research and Practice (*J Cancer Surv*), Journal of Pain and Symptom Management (*J Pain Symptom Manage*), Quality of Life Research (*Qual Life Res*), e Health And Quality of Life Outcomes (*Health Qual Life Outcomes*).

Em síntese, considerando-se tanto a quantidade de artigos publicados, a relevância das citações e presença no Top 10 em todos esses ranqueamentos, identificamos como aquelas mais relevantes na comunidade de *QoL* as publicações Psychooncology, Support

Care Cancer, Quality of Life Research (*Qual Life Res*), BMC Cancer, Cancer e Journal of Clinical Oncology (*J. Clin. Oncolol.*).

FIGURA 19 - Comparação entre as posições das publicações conforme apresentadas nas tabelas 1,2 e 3.



4 CONCLUSÃO

A partir do aumento já extensamente comprovado na incidência, e nos custos financeiros direta e indiretamente relacionados ao câncer, o corpo de literatura científica pertinente ao tema vem acompanhando a tendência e apresentando significativo crescimento na última década.

Estudos de bibliometria possuem a capacidade de auxiliar na compreensão da pesquisa científica em determinada área e permitem melhor entendimento do estado da arte, identificação de tendências e direcionamento de políticas públicas. Devido ao já mencionado crescimento importante do volume de publicações, ferramentas e técnicas de big data vêm sendo cada vez mais incorporadas nesse e em outros campos de pesquisa científica.

Através de análises exploratórias, demonstramos o aumento na relevância do tema qualidade de vida em pesquisas em oncologia, e identificamos também tratar-se de tema interdisciplinar e de amplo interesse na comunidade científica.

O uso de técnicas de processamento de linguagem natural aplicadas à base de dados relativa às publicações relacionadas à qualidade de vida em oncologia nos permitiu identificar a China como principal identidade geográfica, embora Europa lidere como principais identidades cultural e organizacional. Além disso, pudemos criar um método para gerenciamento de artigos científicos, permitindo curadoria para pesquisadores

através da identificação de clusteres e ranqueamento de artigos mais influentes em determinada área de pesquisa.

Análise de redes de citações possibilitou a identificação das relações entre os principais autores e periódicos do tema, além da identificação de grupos de pesquisa por análise de coautoria.

Como limitações do presente estudo podemos citar o fato de que não incluímos na base de dados os textos completos dos artigos científicos, e a escolha algo arbitrária por determinadas técnicas nas análises realizadas, o que pode de certa forma ter algum impacto nos resultados encontrados.

Acreditamos, conforme já exposto, que técnicas de machine learning serão utilizadas de forma ainda mais frequente em análises bibliométricas, auxiliando na extração de valor de corpo científico de volume crescente.

REFERÊNCIAS

1. https://gco.iarc.fr/.

2. https://www.who.int/features/factfiles/cancer/en/.

3. Nayak MG, George A, Vidyasagar MS, Mathew S, Nayak S, Nayak BS, et al. Quality of Life among Cancer Patients. Indian J Palliat Care. 2017;23(4):445-50.

4. Cabral BP, da Graca Derengowski Fonseca M, Mota FB. The recent landscape of cancer research worldwide: a bibliometric and network analysis. Oncotarget. 2018;9(55):30474-84.

5. https://pubmed.ncbi.nlm.nih.gov/.

6. https://github.com/massimoaria/pubmedR.

7. Sigla em inglês para Sistema Online de Busca e Análise de Literatura Médica (Medical Literature Analysis and Retrieval System Online) - https://pt.wikipedia.org/wiki/MEDLINE, acesso em 02/07/2020.

8. United States National Library of Medicine - NLM. Esta biblioteca faz parte do National Institute o Health - NHL. https://en.wikipedia.org/wiki/United_States_National_Library_of_Medicine, acesso em 02/07/2020.

9. https://clarivate.com/webofsciencegroup/solutions/journal-citation-reports/.

10. https://www.scimagojr.com/.

11. MeSH = Medical Subject Headings. É um vocabulário controlado, Thesaurus, usado para catalogar, indexar e buscar informações biomédicas e relacionadas a saúde. https://www.nlm.nih.gov/mesh/meshhome.html, consultado em 03/07/2020.

12. Aksnes DW. Characteristics of highly cited papers. Quality assessment. 2003;12(3):159-170.

13. Lutz Bornmann, Loet Leydesdorff, Rüdiger Mutz, The use of percentiles and percentile rank classes in the analysis of bibliometric data: Opportunities and limits. Journal of Informetrics, 7 (1) (2013), pp. 158-165.

14. Barabási, A., Song, C. & Wang, D. Handful of papers dominates citation. Nature 491, 40 (2012).

15. Wang J, Deng H, Liu B, et al. Systematic Evaluation of Research Progress on Natural Language Processing in Medicine Over the Past 20 Years: Bibliometric Study on PubMed. J Med Internet Res. 2020;22(1):e16816.

16. https://towardsdatascience.com/stop-words-in-nlp-5b248dadad47.

17. https://allenai.github.io/scispacy/. Acesso em 2/07/2020.

18. https://spacy.io/api/annotation - named-entities. Acesso em 2/07/2020.

19. Giller, Graham L., The Statistical Properties of Random Bitstreams and the Sampling Distribution of Cosine Similarity (October 25, 2012). .

20. Mahout in Action de Sean Owen, Robin Anil, Ted Dunning, Ellen Friedman (2012), Capítulo 4, pg. 52.

21. Kuraria, Amit & Jharbade, Nitin & Soni, Manish. (2018). Centroid Selection Process Using WCSS and Elbow Method for K-Mean Clustering Algorithm in Data Mining. International Journal of Scientific Research in Science, Engineering and Technology. 190-195.

22. https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html. Acesso em 02/07/2020.

23. Daniel, T., Casenave, F., Akkari, N. et al. Model order reduction assisted by deep neural networks (ROM-net). Adv. Model. and Simul. in Eng. Sci. 7, 16 (2020).

24. Salman, Raied & Kecman, Vojislav & Li, Qi & Strack, Robert & Test, Erik. (2011). Fast K-Means Algorithm Clustering. International Journal of Computer Networks & Communications.

25. Garfield, E. (1979). Is citation analysis a legitimate evaluation tool? Scientometrics, 1(4), 359–375.

26. Francesco Alessandro Massucci, Domingo Docampo, Measuring the academic reputation through citation networks via PageRank, Journal of Informetrics, Volume 13, Issue 1 (2019), pp. 185-201.

27. Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems, 30(1), 107–117,proceedings of the Seventh International World Wide Web Conference.

28. Becoming Metric-Wise. A Bibliometric Guide for Researchers, Ronald Rousseau, Leo Egghe and Raf Guns, Elsevier/Chandos Publishing (2018). 385 pages.