

Original Paper

# Using Latent Dirichlet Allocation Topic Modeling to Uncover Latent Research Topics and Trends in Renal Cell Carcinoma: Bibliometric Review

Javier De La Hoz-M<sup>1</sup>, PhD; Karime Montes-Escobar<sup>2,3</sup>, PhD; Carlos Alfredo Salas-Macias<sup>4</sup>, PhD; Martha Fors<sup>5</sup>, MD, PhD; Santiago J Ballaz<sup>6</sup>, PhD

<sup>1</sup>Universidad del Magdalena, Santa Marta, Colombia

<sup>2</sup>Departamento de Ciencias Agrícolas, Facultad de Ingeniería Agrícola, Universidad Técnica de Manabí, Portoviejo, Ecuador

<sup>3</sup>Departamento de Formación y Desarrollo Científico en Ingeniería, Facultad de Ingeniería, Ciencia y Tecnología, Universidad Bernardo O'Higgins, Santiago de Chile, Chile

<sup>4</sup>Laboratory of Agroecosystems Functioning and Climate Change – FAGROCLIM, Departamento de Ciencias Agronómicas, Facultad de Ingenierías Agroambientales, Universidad Técnica de Manabí, Santa Ana, Ecuador

<sup>5</sup>Facultad de Ciencias de la Salud, Escuela de Medicina, Universidad de Las Américas, Quito, Ecuador

<sup>6</sup>Facultad de Ciencias de la Salud “Dr. Enrique Ortega Moreira”, Universidad Espíritu Santo, Samborondón, Ecuador

## Corresponding Author:

Martha Fors, MD, PhD

Facultad de Ciencias de la Salud, Escuela de Medicina

Universidad de Las Américas

Granados vía a Nayón sin número

Quito 177022

Ecuador

Phone: 593 95016627

Email: [martha.fors@udla.edu.ec](mailto:martha.fors@udla.edu.ec)

## Abstract

**Background:** Renal cell carcinoma (RCC) is a common, often lethal kidney cancer that originates in the renal cortex. Its incidence is rising, and major factors include smoking, obesity, and hypertension, though its etiology is uncertain. While surgery is effective for localized RCC, treatments for metastatic RCC have advanced significantly due to better diagnostic, prognostic, and predictive tools. Despite this progress, challenges remain, including long-term drug resistance and the complexity of RCC as a diverse group of diseases rather than a single entity.

**Objective:** The aim of this bibliometric review was a comprehensive analysis of the topics and trends in RCC research, offering a foundation for future investigations.

**Methods:** We used R “Bibliometrix” to conduct a bibliographic search in Scopus and PubMed covering publications from 1975 to 2023 to statistically assess the distribution of publications associated with RCC by year, journal, and country. Topic modeling of RCC research was conducted using latent Dirichlet allocation, a Bayesian network-based probabilistic algorithm that identifies unobserved thematic clusters in a collection of text documents. Trends in the retrieved themes were then characterized by using regression slopes over time, across countries, and in different journals. These trends were visualized as a heatmap, which was then used for hierarchical clustering to group similar topics based on their correlation strengths.

**Results:** A total of 35,228 documents from 3070 sources were found, with a steady yearly growth of 9.86% and 118 participating countries. Thirty topics with the best coherence score were found in 8 crucial domains: treatment and therapies, biomolecular and genetic characteristics, disease characteristics and progression, diagnosis and evaluation, metastasis and dissemination, epidemiology and risk factors, related conditions, and pathological features. The pertinent clustergrams that resulted from the heatmaps mirrored the latent Dirichlet allocation’s algorithm identification of major RCC research subjects.

**Conclusions:** Over 50 years, RCC research’s focus has shifted from diagnosis and assessment to a more thorough understanding of disease characteristics and progression. Because many patients are diagnosed with abdominal imaging studies, an emerging topic in RCC is diagnostic imaging and radiological evolution. The advances in omics technologies and the function of microRNA signature in the progression, diagnosis, therapy targeting, and prognosis of RCC have garnered a lot of attention. The discovery of the genetic background has enhanced our understanding of the growth of RCC. Drug resistance, local

RCC ablation, and postoperative surveillance of RCC recurrence following nephrectomy are key future research avenues. The next generation of drug-targeted therapy and immunotherapy will make it possible to successfully treat metastatic RCC following nephrectomy. Neglected topics include the association between ferroptosis and RCC, the long-term assessment of novel treatments, and the application of artificial intelligence on RCC. Our bibliographic review delivered pertinent data for clinical decision-making and the planning of future RCC research.

*JMIR Cancer* 2026;12:e78797; doi: [10.2196/78797](https://doi.org/10.2196/78797)

**Keywords:** bibliometrics; carcinoma renal cell; knowledge discovery; machine learning; research

## Introduction

Renal cell carcinoma (RCC) is the term coined to describe the malignant transformation of proximal renal tubular epithelium within the renal cortex [1]. RCC accounts for approximately 90% of all renal malignancies [2] and for approximately 2% of all cancer diagnoses and cancer deaths worldwide [3]. With a 2:1 ratio of new diagnoses, men are more likely than women to be affected by RCC, whose incidence rises significantly with age. RCC is a serious health problem since it is often a lethal kidney cancer with an increasing incidence worldwide. Its relevance is also tied to challenges in early diagnosis and the development of aggressive subtypes like those involving tumor thrombus. RCC's etiology is uncertain. The 3 main risk factors for RCC include being overweight, having hypertension, and smoking cigarettes [3]. Medical disorders such as chronic kidney disease, hemodialysis, kidney transplantation, polycystic kidney disease, and renal stones are additional risk factors for RCC. Numerous dietary, occupational, environmental, and lifestyle factors have also been linked to the development of RCC [4]. Even though the majority of RCCs are sporadic, 3%-5% of all RCC diagnoses occur in patients younger than 46 years, which suggests an underlying RCC form that is inherited [5].

In terms of treatment, surgery is a successful strategy for managing localized RCC, but conventional chemotherapy is ineffective for treating metastatic RCC. Thankfully, during the past 10 years, amazing progress has been made in treating metastatic RCC, resulting in a significant drop in the cancer's death rates despite a continuous rise in the number of individuals receiving a diagnosis. The main factor for the improvement in RCC survival over the past few decades has been the wide diagnostic, prognostic, and predictive methodologies that are currently available [6,7]. Despite these advancements, long-term RCC drug resistance is still a problem [8]. While accelerating RCC cures is critically needed, research on RCC faces various obstacles. Prior to the discovery of the *VHL* gene, kidney cancer was regarded as a single disease [9]. In fact, kidney cancer is a multitude of diverse diseases, each with its own genetic makeup [10]. This has delayed the search for a cure by impeding reproducibility across studies and appropriate interpretation of the research [11].

RCC offers an intricate and challenging research landscape that hinders scientific progress. The goal of this study was

to conduct a comprehensive and up-to-date summary of the topic structure, novel research avenues, study trends, and knowledge gaps in RCC research. Conventional bibliometric methods, scientific mappings, and network visualization studies fall short in offering this kind of text analysis [12], since they frequently necessitate manual categorization or extensive, subjective human intervention [13-16]. Instead, we sought a corpus of text-based data for research trends and topics by using a topic algorithm model called latent Dirichlet allocation (LDA) [17]. We intended for RCC research documents "to tell the story themselves" and for topics to emerge on their own, without human intervention, and only based on their statistical characteristics.

## Methods

### Search Strategy and Data Collection

This study was based on data obtained from PubMed [18] and Scopus [19] as of April 6, 2023. We chose Scopus over other databases because of its superior coverage in the health sciences field, accurate indexing, and "federated search interface" (ie, functionality), which enables us to query the content found across its sources using a common or standardized search form. The authoritative and comprehensive PubMed is the top choice for searching medical and health sciences literature. Medical Subject Headings (MeSH) terms were used for the PubMed search to improve comprehensiveness. Raw data were stored in *TXT* and *CSV* files, respectively. The R "Bibliometrix" tool of the R Statistical software [20] was used to clean data and integrate the 2 databases' unique publications into a combined dataset (with and without assigned DOI) [18,19,21-23]. At this point, duplicate documents with an assigned DOI (1809 in total) were eliminated. The inclusion criteria for the outcome were all the research documents, written in English in peer-reviewed journals, that were published between 1974 (the earliest article we found) and 2023, and that dealt with RCC. Books, book chapters, gray literature, and reports were not included to avoid noise. The search strings, which used a Boolean computation, are indicated in Table 1. Only those articles with the term "renal cell carcinoma" in their title or abstract were selected. The leader and the other authors reviewed the complete list of all possible acceptable publications. Their reliability and value to the field were based on criteria like the journal's impact factor, author affiliations, and citation count.

**Table 1.** Enhanced information retrieval for research on renal cell carcinoma.

Database	Search data	Search string	Results, n
PubMed	April 6, 2024	“renal cell carcinoma”[Title/Abstract] AND “english”[Language] AND “journal article”[Publication Type] AND 1974/01/01:2023/12/31[Date - Publica- tion]	38,577
Scopus	April 6, 2024	TITLE-ABS ( “renal cell carcinoma” ) AND PUBYEAR>1973 AND PUB- YEAR<2024 AND ( LIMIT-TO ( DOCTYPE , “ar” ) OR LIMIT-TO ( DOCTYPE , “re” ) ) AND ( LIMIT-TO ( SRCTYPE , “j” ) ) AND ( LIMIT-TO ( LANGUAGE , “English” ) )	40,479

**Bibliometrix Analysis**

Our bibliometric review procedure adhered to best practice guidance published elsewhere [24] (Checklist 1). A preliminary descriptive analysis of the retrieved information was carried out using the R package bibliometrix [22]. This open-source application analyzes publication and citation metrics using mathematical and statistical methods to obtain a broad picture of the scientific output that was within the purview of the study. The questions answered were as follows: (Q1) What are the primary research topics in RCC? (Q2) How have RCC research questions changed over time? (Q3) How are these research topics distributed across countries and scientific journals? Three levels of analysis—countries, sources, and authors—were included at this stage to answer the above questions.

The annual growth rate of publications was calculated using the Bibliometrix package in R, which computes the compound annual growth rate (CAGR) using the following equation:

$$CAGR = (V_f/V_i)^{\frac{1}{n}} - 1,$$

(1)

where  $V_f$  is the number of publications in the final year of the study period,  $V_i$  is the number of publications in the initial year of the study period, and  $n$  is the number of years between the initial and final year.

**Latent Dirichlet Allocation**

The unsupervised machine learning algorithm LDA [17] was applied to identify topics. Considered an extension of the probabilistic latent semantic analysis, it has its roots in Bayesian models [17,25]. Topics in LDA are thought of as multinomial distributions of vocabulary terms, in which each word has a given probability of occurring in a topic. This leads to the prominence of words that are more frequently used in a topic, creating clusters that reflect specific underlying themes. LDA does not require prior knowledge of the topics or the way they are presented in the texts. Rather, topics merely flow from the statistical properties of the data and the model’s underlying assumptions. For thematic analysis, it was decided to use abstracts rather than entire texts because topics are more coherent and ranked higher in

large document collections. Inaccurate or noisy terms have less impact on topic word distribution [26].

The LDA model was validated using the Cv metric, which is grounded in the distributional hypothesis stating that words with similar meanings tend to coexist in similar contexts [27]. In other words, Cv rates the semantic similarity of words within a topic (ie, topic interpretability). The Cv score was calculated by looking at word co-occurrence statistics in a reference corpus and their conformity to human-like semantic interpretation. The package textmineR was used for such analysis, which made it easier to determine the ideal number of topics ( $k$ ) for the study. Because higher scores indicate better interpretability, a model with the highest coherence score among those in the study rank (from  $k=4$  to 50) was selected to reach a balance between granularity and thematic clarity.

**Identifying Research Topics**

The procedure for identifying topics through LDA was divided into 3 stages: (1) preprocessing, (2) construction of the LDA model, and (3) assigning labels to topics. LDAShiny [28], an open-source R package that uses Bayesian inference for LDA and machine learning algorithms to improve the analytical process, was selected for the first 2 stages.

**Preprocessing**

Converting all documents into a standardized format for ease of handling was the objective of the stage known as “Text refining” [29]. Initially, textual data consisted solely of character sets. To enhance topic coherence, each abstract underwent tokenization using bigrams, which are consecutive unigram combinations. This process involves converting text to lowercase and removing punctuation marks, dashes, brackets, numbers, spaces, and “stop words.” The list of stop words was extracted from standard libraries such as Natural Language ToolKit and Snowball and was modified to include unrelated terms unique to the medical and technical domains.

The preprocessed data result in the creation of a document-term matrix in which each document is represented as a vector containing an unordered collection of words. If the corpus contains a total of  $V$  words, each document becomes a  $V$ -dimensional vector, with the value of each element representing the frequency of the corresponding word in the document.

## Construction of the LDA Model

LDA assumes that topics are shared by all documents in the collection, while subject proportions vary stochastically between documents, as they are randomly extracted from a Dirichlet distribution [30]. Establishing the expected number of topics was done a priori, making it a nontrivial task to choose the right number of topics ( $k$ ) for a given collection of items. Since the optimal number of topics was unknown beforehand, we generated different models ranging from 4 to 40 topics. We ran 1000 iterations for Gibbs sampling [31] and utilized the default values of the LDAShiny package for Dirichlet parameters  $\alpha$  and  $\beta$ . We used Cv as the topic coherence measure of the topics generated by LDA models [32].

## Assigning Labels to Topics

The LDA model generates topics without semantic labels. Given that algorithmic analyses are not always able to fully capture the implicit meanings of human language, manual labeling is widely recognized as a normal practice in topic modeling [32]. The manual topic labeling involved a diverse team of 7 experts, including the authors and independent scholars with backgrounds in oncology research and bibliometric analysis. Team members were provided with 2 sources of information: the lists of most frequently occurring words (presumably) provided by the model and a sample of 3 document titles with their corresponding summaries classified by the algorithm. They were asked to independently validate and summarize the identified topics and investigate existing literature to identify research trends, gaps, and influential works. Discrepancies between team members were resolved through remote communications to minimize bias, improve rigor, and ensure the theme structures and semantic interpretation were aligned with the research objective. Reference [32] provides a guide to the procedures used to ensure the trustworthiness of the labels issued, with the difference that we used up to 16 annotators rating candidate labels.

By providing the most relevant and thematically aligned examples within each topic, these articles guaranteed readability and clarity. The 2 articles chosen allowed for a concise and efficient summary of each topic's key ideas without overwhelming the analysis and the reader. This approach strikes a balance between realistic representation and practical interpretability. These articles were then condensed into succinct summaries that encapsulated the essence of each topic. This manual approach not only provided a gold-standard reference [32] but also ensured interpretability and utility in the context of the study.

## Quantitative Indices

For each topic, additional characteristics were revealed, especially at the journal and country levels, through statistical description based on the probability distributions of document-topic and topic-word acquired through LDA. To make results and findings more evident, we used certain quantitative indices suggested by Xiong and colleagues [33],

which were obtained by adding document-topic and topic-word distributions. The indexes were described as follows.

The distribution of topics over time was obtained by the following equation:

$$\theta_k^y = \sum_{m \in y} \theta_{mk} / n^y, \quad (2)$$

where  $m \in y$  represents articles published each year,  $\theta_{mk}$  is the proportion of the  $k$ th topic in each item, and  $n^y$  is the total number of articles published in the year.

Topic distribution across journals was defined as the ratio of the  $k$ th topic in the journal  $j$ :  $\theta_k^j$  as indicated in the following equation:

$$\theta_k^j = \sum_{m \in j} \theta_{mk} / n^j, \quad (3)$$

where  $m \in j$  represents the articles in a particular journal,  $\theta_{mk}$  is the proportion of the  $k$ th topic on each item, and  $n^j$  is the total number of articles published in the journal  $j$ .

Topic distribution across countries was defined as the ratio of the  $k$ th topic in the country  $c$ , as in the following equation:

$$\theta_k^c = \sum_{m \in c} \theta_{mk} / n^c, \quad (4)$$

where  $m \in c$  represents the articles in a particular country,  $\theta_{mk}$  is the proportion of the  $k$ th topic on each item, and  $n^c$  is the total number of articles published in the country  $c$ .

## Statistics

With the purpose of facilitating the characterization of the topics in terms of their tendency, topic datasets were also submitted to simple regression slopes where the year, country, and journal were the dependent variables, while the proportion of the topics in the corresponding year, country, and journal was the response variable [34]. From the regression slopes, we determined the directionality of these trends and set a significance threshold of  $P < .01$  (Eq2). Topics that showed statistically significant positive slopes were identified as having upward trends, while those with statistically significant negative slopes were in decline. Tendencies were finally visualized using the *ggcorrplot* library of R to represent correlation strengths as a heatmap matrix. The color-mapped matrix was subjected to advanced hierarchical clustering analysis in order to investigate and compile correlation datasets, given its visual form like a tree-shaped dendrogram. The *Agnes* function with *Ward's* method showed the agglomerative hierarchical clustering of variables. Each leaf of the dendrogram corresponded to one observation (variable), and the fusion height showed the dissimilarity between 2 observations on the vertical axis. A cut height for cluster identification was calculated using the Average Silhouette method [35].

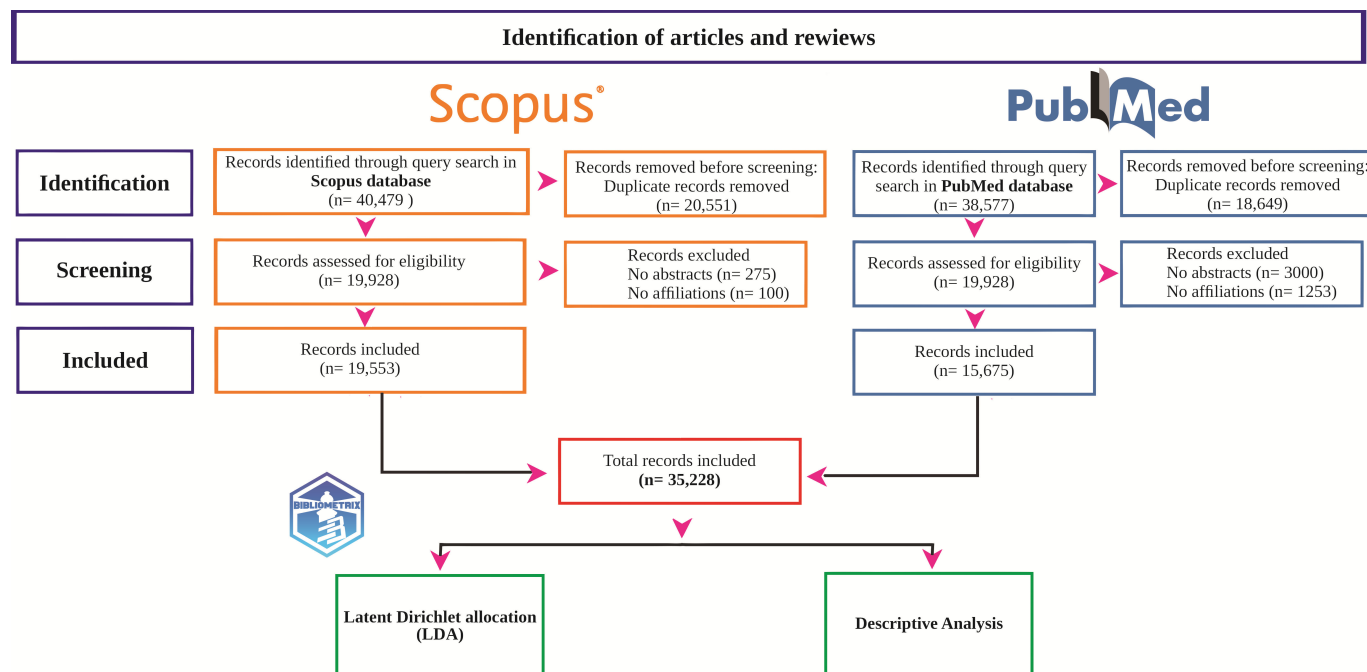
## Results

### Overview of the Dataset

The consolidated dataset was obtained by combining the results and removing duplicates, totaling 39,856 articles (Figure 1). After the relevant Excel file was created, 4628 articles lacking titles, abstracts, or affiliations were removed. Within this extensive dataset, 35,228 documents were assembled, demonstrating an annual growth rate of 9.86%.

The extensive summary of the key descriptive characteristics pertaining to RCC from 1974 to 2023 can be found in Table 2. A substantial number of information sources, 3070 in total, were revealed by the data. Given that the document's average age was 11.6 years (the time since the publication of the examined articles), it is likely that much of the research was carried out some time ago. On average, however, each document received 32.35 citations, demonstrating their influence and recognition in the field (Table 2).

**Figure 1.** The workflow for article selection and bibliometric analysis in renal cell carcinoma using the PubMed and Scopus databases.



**Table 2.** Comprehensive overview of key descriptive characteristics and publication metrics on renal cell carcinoma from 1974 to 2023. Retrieved from the PubMed and Scopus databases.

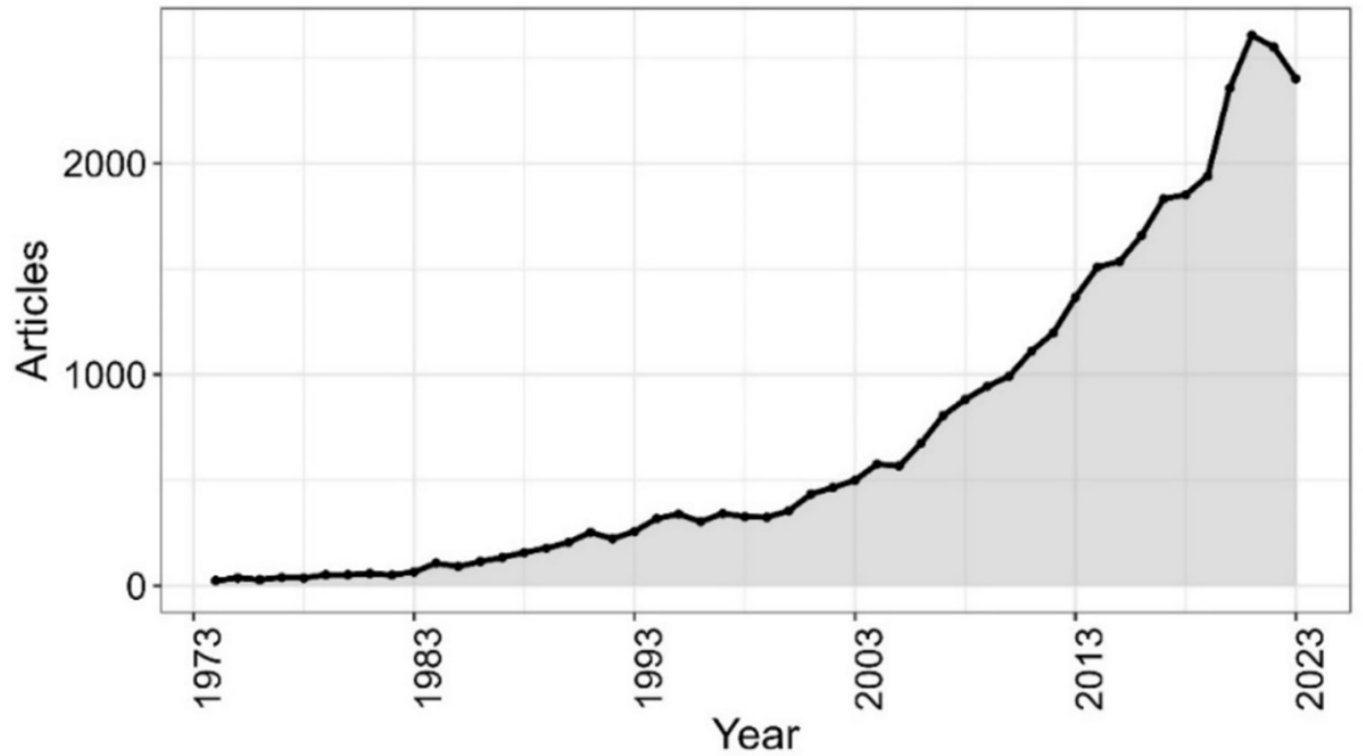
Description	Results
Main information about data	
Timespan, range	1974-2023
Sources (journals, books, etc), n	3070
Total documents, n	35,228
Annual growth rate (%), mean	9.86
Document age (y), mean	11.6
Citations per document, mean	32.35
Document contents, n	
Keywords plus	52,792
Author's keywords	31,248
Authors, n	
Total authors	95,238
Authors of single-authored documents	608
Author collaboration	
Single-authored docs, n	769
Coauthors per doc, mean	7.55



Description	Results
Document types	
Original research	30,913
Review	4315

With respect to the yearly output of documents on RCC, [Figure 2](#) provides an overview of increased production from 1974 (24 articles) to 2023 (2401 articles). Throughout the 1980s and 1990s, the document production grew modestly. However, the 2000s saw a significant increase in production, reaching a peak in 2017 (1832 articles) and indicating a solid trend in recent years. [Table 3](#) lists the top 30 scientific journals, while a global map ([Figure 3](#)) shows the 118 countries that were involved in RCC research. With 10,308 publications, the United States clearly was the top contributor.

**Figure 2.** Annual production of documents on renal cell carcinoma from 1974 to 2023.

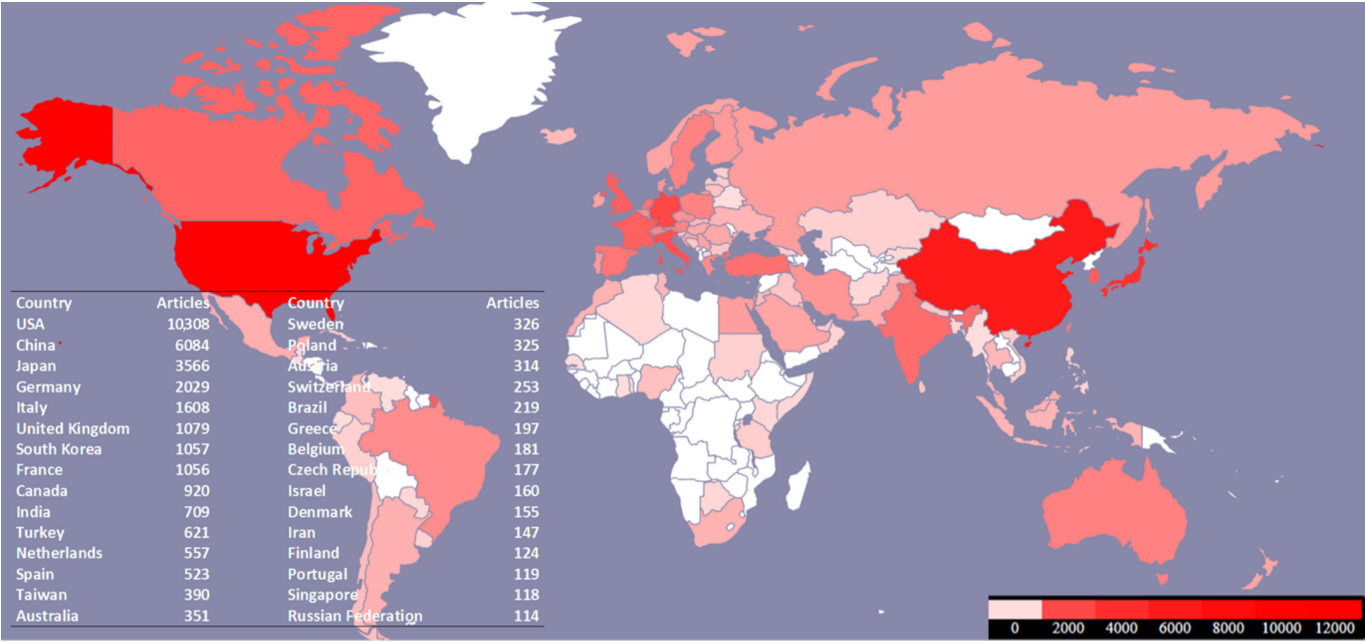


**Table 3.** Top 30 scientific journals for research on renal cell carcinoma based on articles published between 1974 and 2023.

Source	Abbreviation	Articles, n
<i>Journal of Urology</i>	<i>J Urol</i>	1161
<i>Urology</i>	<i>Urology</i>	839
<i>Urologic Oncology: Seminars and Original Investigations</i>	<i>Urol Oncol</i>	560
<i>European Urology</i>	<i>Eur Urol</i>	542
<i>Frontiers in Oncology</i>	<i>Front Oncol</i>	481
<i>Cancer</i>	<i>Cancer</i>	478
<i>Clinical Genitourinary Cancer</i>	<i>Clin Genitourin Cancer</i>	474
<i>BJU International</i>	<i>BJU Int</i>	451
<i>International Journal of Urology</i>	<i>Int J Urol</i>	422
<i>Oncotarget</i>	<i>Oncotarget</i>	363
<i>PLOS One</i>	<i>PLoS One</i>	362
<i>Cancers</i>	<i>Cancers</i>	350
<i>Urologia Internationalis</i>	<i>Urol Int</i>	316

Source	Abbreviation	Articles, n
<i>International Journal of Cancer</i>	<i>Int J Cancer</i>	311
<i>Clinical Cancer Research</i>	<i>Clin Cancer Res</i>	308
<i>Oncology Letters</i>	<i>Oncol Lett</i>	307
<i>World Journal of Urology</i>	<i>World J Urol</i>	297
<i>British Journal of Cancer</i>	<i>Brit J Cancer</i>	291
<i>BMC Cancer</i>	<i>BMC Cancer</i>	273
<i>Scientific Reports</i>	<i>Sci Rep</i>	248
<i>American Journal of Surgical Pathology</i>	<i>Am J Surg Pathol</i>	236
<i>American Journal of Roentgenology</i>	<i>Am J Roentgenol</i>	210
<i>Journal of Clinical Oncology</i>	<i>J Clin Oncol</i>	206
<i>Cancer Research</i>	<i>Cancer Res</i>	200
<i>Human Pathology</i>	<i>Hum Pathol</i>	196
<i>Oncology Reports</i>	<i>Oncol Rep</i>	189
<i>International Urology and Nephrology</i>	<i>Int Urol Nephrol</i>	188
<i>International Journal of Molecular Sciences</i>	<i>Int J Mol Sci</i>	186
<i>Medicine (United States)</i>	<i>Med (United States)</i>	172
<i>Journal of Endourology</i>	<i>J Endourol</i>	165

**Figure 3.** Distribution of geographical origins in the analysis of 35,228 published articles on renal cell carcinoma from 1974 to 2023. The table displays the top 30 countries with the highest research production.



**Latent Dirichlet Allocation**

The methodological rigor, reproducibility, and accuracy of LDA were ensured using advanced bibliometric tools such as Bibliometrix and Textminer. There were 30 topics with the

best coherence score in the LDA model (Figure 4). The terms with the highest probabilities and semantically relevant labels for each latent topic are shown in Table 4.

**Table 4.** Topics discovered from 35,228 articles on renal cell carcinoma published between 1974 and 2023.

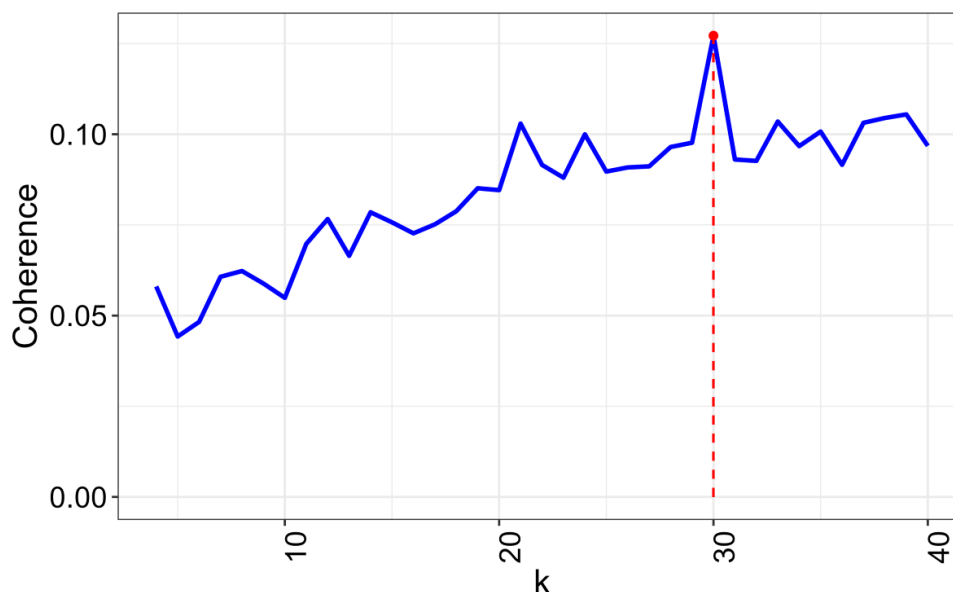
Terms	Article numbers, n	Top terms	Label	Themes	Prevalence (%)
t_1	1411	inhibitor, sunitinib, treatment, target, drug, kinas, vegf, growth, factor, tki, sorafenib, resist, agent, tyrosin, growth_factor	Inhibitors in treatment	Treatment and therapies	3.455

Terms	Article numbers, n	Top terms	Label	Themes	Prevalence (%)
t_2	563	rcc, level, patient, serum, control, rcc_patient, increas, group, concentr, elev, blood, distribut, plasma, healthi, compar	Serum levels and patient control	Diagnosis and evaluation	2.414
t_3	1240	express, tissu, normal, protein, tumor, level, correl, mrna, sampl, kidney, posit, compar, marker, normal_tissu, express_level	Protein expression in tissues and tumors	Biomolecular and genetic characteristics	3.866
t_4	1370	gene, mutat, tumor, chromosom, genet, dna, loss, alter, famili, sequenc, region, identifi, methyl, variant, r	Genetic mutations and chromosomal alterations	Biomolecular and genetic characteristics	3.067
t_5	2113	rcc, clear, tumor, papillari, type, case, subtyp, carcinoma, featur, clear_rcc, prcc, posit, chromophob, histolog, pattern	Histological features and subtypes	Disease characteristics and progression	4.537
t_6	1293	tumor, surgic, thrombu, complic, oper, nephrectomi, surgeri, resect, laparoscop, blood, vena, ivc, postop, perform, approach	Surgical approaches and complications	Pathological features	3.053
t_7	939	vhl, hif, hif_factor, hif_1, hif_2, hif_3, hif_4, hif_5, hif_6, hif_7, hif_8, hif_9, hif_10, hif_11, hif_12, hif_13, hif_14, hif_15, hif_16, hif_17, hif_18, hif_19, hif_20, hif_21, hif_22, hif_23, hif_24, hif_25, hif_26, hif_27, hif_28, hif_29, hif_30, hif_31, hif_32, hif_33, hif_34, hif_35, hif_36, hif_37, hif_38, hif_39, hif_40, hif_41, hif_42, hif_43, hif_44, hif_45, hif_46, hif_47, hif_48, hif_49, hif_50, hif_51, hif_52, hif_53, hif_54, hif_55, hif_56, hif_57, hif_58, hif_59, hif_60, hif_61, hif_62, hif_63, hif_64, hif_65, hif_66, hif_67, hif_68, hif_69, hif_70, hif_71, hif_72, hif_73, hif_74, hif_75, hif_76, hif_77, hif_78, hif_79, hif_80, hif_81, hif_82, hif_83, hif_84, hif_85, hif_86, hif_87, hif_88, hif_89, hif_90, hif_91, hif_92, hif_93, hif_94, hif_95, hif_96, hif_97, hif_98, hif_99, hif_100	Hypoxia factors and related proteins	Pathological features	2.654
t_8	575	effect, treatment, ablat, control, local, search, evid, review, includ, meta, radiat, perform, systemat, outcom, percutan	Treatment effects and local ablation	Treatment and therapies	2.281
t_9	770	immun, pd, respons, immunotherapi, combin, nivolumab, checkpoint, ici, inhibitor, immun_checkpoint, therapi, checkpoint_inhibitor, treatment, death, anti	Immunotherapy and immune responses	Treatment and therapies	2.416
t_10	544	model, predict, score, risk, base, valid, group, curv, clinic, perform, cohort, featur, characterist, set, auc	Risk prediction models and clinical assessment	Epidemiology and risk factors	2.782
t_11	1593	imag, ct, enhanc, lesion, mass, contrast, evalu, tomographi, mri, comput, detect, phase, comput_tomographi, scan, find	Diagnostic imaging and radiological evaluation	Diagnosis and evaluation	3.633
t_12	1794	patient, month, surviv, median, mrcc, progress, o, group, pf, treat, metastat, free, line, progress_free, receiv	Survival and disease progression	Disease characteristics and progression	4.504
t_13	125	tumor, node, lymph, lymph_node, invas, metastasi, distant, involv, stage, crcc, posit, distant_metastasi, presenc, patient, node_metastasi	Metastasis and lymph node involvement	Metastasis and dissemination	1.776
t_14	724	metastat, metastas, metastasi, primari, patient, bone, site, rcc, lung, lesion, brain, primari_tumor, resect, pet, diseas	Metastasis and lesions in other organs	Metastasis and dissemination	2.724
t_15	324	patient, syndrom, develop, symptom, diseas, clinic, common, earli, relat, occur, sever, adult, hypertens, infect, manifest	Syndromes and clinical manifestations	Pathological features	2.108
t_16	1969	tumor, activ, human, effect, induc, increas, mice, antibodi, antigen, anti, deriv, specif, growth, line, cytotox	Tumor activity and immune responses	Disease characteristics and progression	4.218



Terms	Article numbers, n	Top terms	Label	Themes	Prevalence (%)
t_17	2190	patient, nephrectomi, year, recurr, group, surgeri, follow, rate, rang, underw, local, radic, month, partial, diseas	Nephrectomy and recurrence	Metastasis and dissemination	5.242
t_18	1147	rcc, risk, ci, increas, associ, ag, patient, incid, compar, ratio, popul, year, mortal, interv, data	Risk factors and epidemiology	Epidemiology and risk factors	3.858
t_19	2049	patient, respons, treatment, dose, toxic, dai, week, event, advers, receiv, diseas, progress, evalu, phase, efficaci	Toxicity and adverse events in treatments	Pathological features	4.779
t_20	102	bladder, prostat, urolog, urinari, health, urotheli, tsc, prostat_cancer, malign, kluwer, lippincott, wolter, wolter_kluwer, william, wilkin	Urological cancers and related conditions	Related conditions	1.628
t_21	351	kidnei, diseas, long, term, transplant, function, long_term, develop, chronic, kidnei_diseas, egfr, diabet, dialysi, donor, recipi	Chronic kidney disease and transplant	Related conditions	1.967
t_22	1917	ccrcc, gene, clear, identifi, express, relat, prognosi, clear_ccrcc, ccrcc_patient, cancer, pathwai, biomark, potenti, data, genom	Gene expression and prognosis	Biomolecular and genetic characteristics	4.362
t_23	487	develop, molecular, potenti, recent, therapeut, provid, approach, clinic, understand, import, review, research, base, applic, strategi	Molecular advances and therapeutics	Pathological features	3.026
t_24	1388	surviv, patient, prognost, factor, independ, multivari, specif, cox, o, prognosi, free, prognost_factor, outcom, predictor, specif_surviv	Prognostic factors and survival	Disease characteristics and progression	4.137
t_25	202	cancer, type, lung, breast, kidnei_cancer, melanoma, cancer_patient, includ, lung_cancer, small, breast_cancer, research, acid, metabol, colorect	Nonrenal cancers and comparative analysis	Pathological features	2.354
t_26	555	malign, tumor, diagnosi, biopsi, benign, pancreat, neoplasm, thyroid, mass, case, diagnost, lesion, diagnos, carcinoma, small	Diagnosis and characterization of tumors	Diagnosis and evaluation	2.559
t_27	3022	rcc, mir, prolifer, inhibit, express, role, regul, assai, line, invas, effect, target, apoptosi, promot, migrat	Gene regulation and microRNA expression	Biomolecular and genetic characteristics	5.728
t_28	604	tumor, stage, grade, size, patholog, pt, tumor_size, low, nuclear, clinic, fuhrman, correl, necrosi, fuhrman_grade, tnm	Staging and pathological features	Pathological features	2.986
t_29	1168	therapi, treatment, clinic, target, improv, trial, advanc, system, review, manag, metastat, rcc, target_therapi, benefit, diseas	Advanced therapies and management of metastatic disease	Treatment and therapies	3.795
t_30	2699	case, report, year, present, rare, reveal, kidnei, left, mass, diagnosi, adren, report_case, examin, diagnos, literatur	Clinical presentation and diagnosis of rare cases	Diagnosis and evaluation	4.089

**Figure 4.** Evaluation of coherence scores for topic models in oncology across different numbers of topics ( $k$ ). Coherence: measures the semantic quality of topics. Higher values indicate stronger word relationships. Topic models: identify thematic patterns in texts; in oncology, they reveal key research areas;  $k$  represents the number of topics in the model. Helps determine the optimal value for thematic analysis.



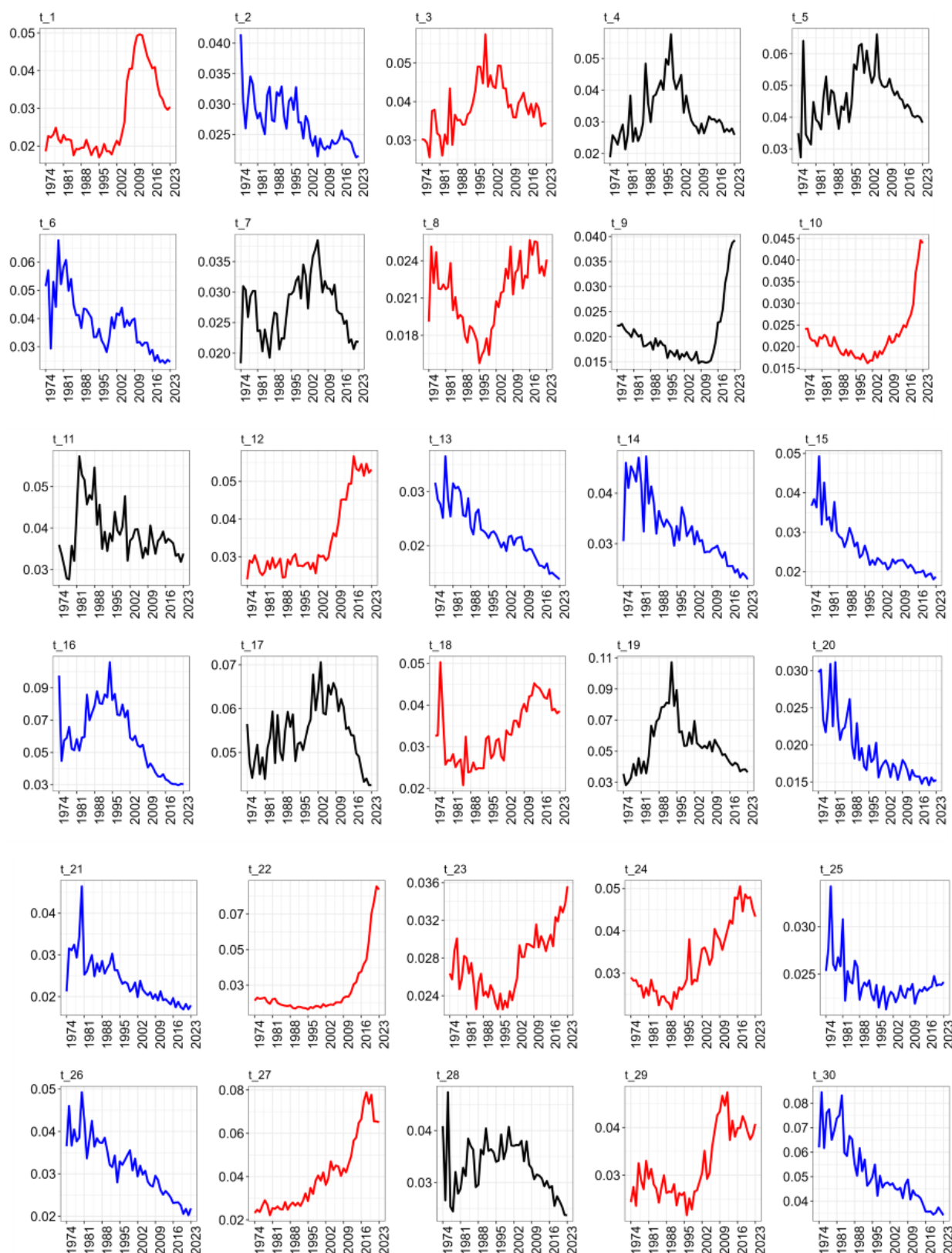
The 30 identified topics were in 8 crucial domains of RCC research:

- **Treatment and Therapies:** This category focused on different approaches to treating RCC, including targeted inhibitors (t\_1), local ablation effects (t\_8), immunotherapy responses (t\_9), and advanced management of metastatic disease (t\_29).
- **Biomolecular and Genetic Characteristics:** Here, the emphasis was on understanding the molecular and genetic makeup of RCC, covering topics like protein expression in tissues and tumors (t\_3, t\_22, and t\_27) and genetic mutations and chromosomal alterations (t\_4).
- **Disease Characteristics and Progression:** This category delved into the histological features and subtypes of RCC (t\_5), as well as the dynamics of tumor activity and immune responses (t\_16), and survival rates and disease progression (t\_12 and t\_24) and pathological features (t\_28).
- **Diagnosis and Evaluation:** Topics in this category included diagnostic imaging and radiological evaluation (t\_2 and t\_11) for RCC detection and the characterization of tumors (t\_26 and t\_30) for accurate diagnosis.
- **Metastasis and Dissemination:** Here, the focus was on understanding how RCC spreads, including its involvement with lymph nodes (t\_13), lesions in other organs (t\_14), and the recurrence of the disease post nephrectomy (t\_17).
- **Epidemiology and Risk Factors:** This category examined the risk prediction models and clinical assessment tools (t\_10) used to evaluate RCC risk, as well as the epidemiological factors associated with the disease (t\_18).
- **Related Conditions:** Topics here explored conditions related to RCC, such as urological cancers (t\_20), chronic kidney disease, and transplant issues (t\_21).
- **Pathological Features:** Finally, this category encompassed various pathological features of RCC, including surgical approaches and complications (t\_6), hypoxia factors and related proteins (t\_7), syndromes and clinical manifestations (t\_15), toxicity and adverse events in treatments (t\_19), molecular advances and therapeutics (t\_23), and comparisons with nonrenal cancers (t\_25).

### Topic Trends

The topic distribution by document  $\theta_m$  was added to compute the average probability  $\theta_k^y$  of all the articles published in a particular year to identify the trends (Figure 5). We found that the probabilities of some topics steadily increased over time (red). Black indicates topics with no discernible trend, whereas blue denotes topics with a decreasing behavior.

**Figure 5.** Trends of research topics in renal cell carcinoma between 1973 and 2023: increasing (red), decreasing (blue), and stable (black) topic dynamics over time.



## Heatmaps

Although the discovery of 30 unique themes with high coherence scores, a granular and nuanced analysis of the research landscape, was made possible by the LDA method,

there was a need for validating the theme output by a form of visual data analytics. Heatmaps helped examine and understand how thematic patterns were related to variables like publication year, country, and journal. Red highlight-

ing indicates the strongest associations among variables, reflecting higher correlation levels.

Figure 6A illustrates correlations between specific topics and years. Red highlighting indicates the strongest associations among variables, reflecting higher correlation levels. For example, within Cluster 4, Topic 16 (t<sub>16</sub>) “Tumor Activity and Immune Responses” is primarily associated with the years 1993, 1985, 1988, 1990, 1992, 1991, 1989, 1995, 1994, 1996, and 1999. Topic 17 (t<sub>17</sub>) “Nephrectomy and Recurrence” exhibits stronger correlations with the years

1993, 1990, 1992, 1991, 1989, 1995, and 1994. In Cluster 3, Topic 22 (t<sub>22</sub>) “Gene Expression and Prognosis” significantly correlates with the years 2020, 2021, 2022, and 2023, while Topic 27 (t<sub>27</sub>), also addressing “Gene Regulation and microRNA Expression,” shows significant associations with the years 2017, 2018, 2019, and 2020. Finally, in group 3, Topic Thirty30 (t<sub>30</sub>) “Clinical Presentation and Diagnosis of Rare Cases” is linked with the years 1982, 1983, 1981, 1978, 1977, and 1975.

**Figure 6.** Heatmaps to correlate topics with year (A), country (B), and source (C).

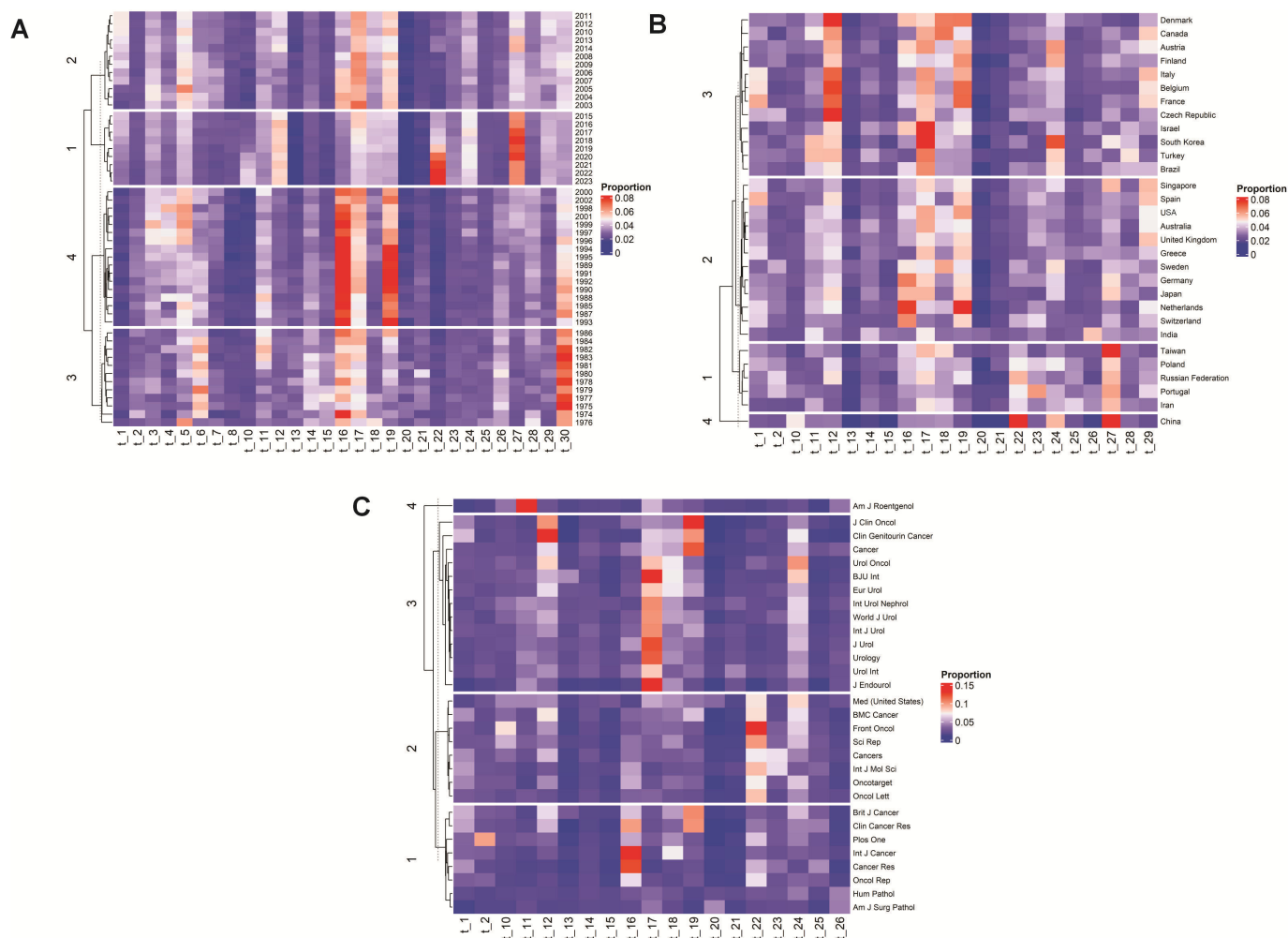


Figure 6B presents the interactions between countries and topics across different groups. In group 1, Topic 27 (t<sub>27</sub>), “Gene Regulation and microRNA Expression,” is predominantly correlated with Taiwan. Group 2’s topics, “Tumor Activity and Immune Responses” and “Toxicity and Adverse Events in Treatments,” are closely associated with the Netherlands. For group 3, “Survival and Disease Progression” shows significant correlations with the Czech Republic, Denmark, Belgium, France, and Italy, and “Nephrectomy and Recurrence” is notably linked with Israel and South Korea. Lastly, in group 4, “Gene Expression and Prognosis” and “Gene Regulation and microRNA Expression” are primarily connected with China.

Figure 6C outlines the dominant publication patterns for specific topics within various journals. In group 1, “Tumor Activity and Immune Responses” is predominantly linked with the *International Journal of Cancer* and *Cancer Research*. Group 2’s topic, “Gene Expression and Prognosis,” has a significant association with *Frontiers in Oncology*. In group 3, “Survival and Disease Progression” relates to *Clinical Genitourinary Cancer*, and “Nephrectomy and Recurrence” is associated with *BJU International*, *Journal of Endourology*, and *Urology*. Lastly, in group 4, “Diagnostic Imaging and Radiological Evaluation” is prominently linked to the *American Journal of Roentgenology*. LDA algorithm’s identification of the key RCC research themes was mirrored

in the relevant clustergrams that emerged from the hierarchical cluster analysis of the heatmaps.

## Discussion

### Principal Findings

By displaying each topic as a group of related words, LDA was able to identify latent (hidden) topics within a corpus of documents and demonstrate how each document may be represented as a combination of these topics. According to this method, RCC research has evolved over the past 50 years from concentrating on surgery to comprehending its genetic underpinnings and the impact of new treatments like immune checkpoint inhibitors and targeted therapies, which have improved the prognosis of metastatic RCC.

The disease characterization, progression, and pathological features have dominated the RCC research landscape for the past 50 years (t<sub>16</sub> and t<sub>19</sub>) [10]. Kidney cancer was regarded as a single disease until the *VHL* gene was discovered [9]. Since then, scientists have realized that kidney cancer is a multitude of diverse diseases, each with its own genetic makeup. Although the characterization of genetic mutations and chromosomal alterations linked to the growth of RCC [36,37] has recently improved our knowledge of this kidney cancer (t<sub>22</sub>), RCC research does emphasize the necessity of assessment of treatment effects' safety and efficacy, as well as surveillance of RCC recurrence following nephrectomy. Therefore, a better understanding of resistance mechanisms, molecular prognosis (t<sub>22</sub>), and immunological responses (t<sub>16</sub>) is essential.

The advances in omics technologies over the last 10 years constitute a promising area of personalized RCC cure [1,38,39]. In addition, the microRNA signature in RCC and its function in progression, diagnosis, therapy targeting, and prognosis of RCC (t<sub>27</sub>) have also received particular attention [40,41]. Recent RCC research also moves toward earlier cancer detection through broad imaging and radiological evolution (t<sub>11</sub>). RCC has a difficult pathological classification, since the histological analysis reveals three most recurrent sporadic types: clear-cell RCC (70%-75%), papillary RCC (10%-15%), and chromophobe RCC (5%) [42]. It is predicated on morphologies, architecture, underlying genetic abnormalities, and tumoral protein expression [43,44]. Pathologists can detect these malignancies more accurately and provide better treatment plans and patient outcomes if they are aware of the significance of these markers (t<sub>12</sub> and t<sub>19</sub>).

A major cause for worry is the postoperative surveillance of RCC recurrence following nephrectomy (t<sub>17</sub>) [45]. Research on local RCC ablation and assessment of treatment effects on safety and efficacy is still ongoing (t<sub>12</sub>, t<sub>16</sub>, and t<sub>19</sub>). The discovery of tailored medication like tyrosine kinase inhibitors (also called TKIs), such as Sunitinib and Sorafenib, has been beneficial in treating metastatic RCC [46, 47]. Another well-established component of RCC treatment is checkpoint inhibitor immunotherapy [48], which shows its superior therapeutic efficacy when combined with TKIs [49].

The next generation of TKIs and immunotherapy (t<sub>16</sub>) [50] is being developed to overcome some unfavorable outcomes with kinase inhibitors [51,52] and immunotherapy (t<sub>19</sub>) [53], as well as the emergence of drug resistance [54]. Only a small number of themes still earn little attention. The topics "Serum Levels and Patient Control" (t<sub>2</sub>), "Hypoxia Factors and Related Proteins" (t<sub>7</sub>), and "Risk Prediction Models and Clinical Assessment" (t<sub>18</sub>) are overlooked and should require additional attention.

In line with the LDA analysis and our own scientific expectations and goals, it was also feasible to identify untapped topics that have not yet been covered by scholarly literature. Intriguing lines of inquiry are the prognostic value of vascular endothelial growth factor [55], endostatin [56], C-reactive protein [57], the hypoxia-induced pathway [58], and ferroptosis [59]. The involvement of chronic inflammation [60] and gut and urinary microbiota in immune modulation of metastatic RCC [61] remains poorly investigated. Clinical judgments and patient stratification in RCC may be enhanced by the creation of new predictive models [62] based on, for instance, genetic biomarkers [63]. The application of artificial intelligence is another potential topic that could help physicians in identifying RCC subtypes by analyzing computed tomography scans, as well as in deconstructing complex epidemiological and environmental factors that influence RCC occurrence, like hypoxia [64].

In contrast to traditional bibliometric analysis, the LDA approach effectively extracted potential themes and inferred implicit information from a large collection of documents. The LDA approach and other topic modeling methods like co-citation and keyword co-occurrence cannot be compared with the same conceptual granularity or depth. LDA goes beyond who cites whom (ie, intellectual connection and research lineage) and uncovers the underlying conceptual themes that bind the literature, which may not be immediately apparent from citation patterns alone. Unlike a list of keywords, which reveals basic relationships, LDA organizes these co-occurring words into meaningful higher-order themes, offering a more detailed knowledge of topic relationships and structure. For this reason, the LDA analysis was not affected by the potential simplicity of the keywords chosen in the study.

The subjectivity involved in manually labeling LDA topics, the possibility of missing publications by using only 2 databases, the linguistic bias introduced by excluding articles written in languages other than English, the constraints of the "bag-of-words" model which disregards grammar and context, and the effects of excluding literature like book chapters are some limitations of the study. Despite the constraints, LDA can disclose "unknown unknowns" by revealing unarticulated or unacknowledged themes. It can also give an overview of the research landscape to identify new topics and interdisciplinary connections, as well as demonstrate how old themes are resurfacing in new ones. As a result, LDA remains the most often used natural language model [13,14,65].



## Conclusions

This review offered a thorough summary of how research on RCC has changed over the previous 50 years. LDA helped identify important emerging trends in treatment development to address drug resistance and undesirable side effects, surgical techniques, and immunotherapy advancements,

among other topics pertinent to clinical practice and medical research. In summary, this study presents a methodological synthesis of the development of RCC research and delivers pertinent data for clinical decision-making, early identification, and the planning of new biomedical research.

## Funding

No external financial support or grants were received from any public, commercial, or not-for-profit entities for the research, authorship, or publication of this article.

## Data Availability

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

## Authors' Contributions

Conceptualization: JDLH-M, KM-E

Data curation: KM-E, JDLH-M

Formal analysis: JDLH-M, KM-E, CAS-M

Investigation: JDLH-M, KM-E, CAS-M, MF, SJB

Methodology: JDLH-M, KM-E, CAS-M

Supervision: MF, SJB

Validation: JDLH-M, KM-E, CAS-M, MF, SJB

Writing – original draft: KM-E, SJB

Writing – review & editing: MF, SJB

All the authors equally contributed to the writing of the final version of the manuscript and were responsible for its content.

## Conflicts of Interest

None declared.

## Checklist 1

Bibliometric analysis checklist.

[[PDF File \(Adobe File\), 97 KB-Checklist 1](#)]

## References

1. Hsieh JJ, Purdue MP, Signoretti S, et al. Renal cell carcinoma. *Nat Rev Dis Primers*. Mar 9, 2017;3:17009. [doi: [10.1038/nrdp.2017.9](#)] [Medline: [28276433](#)]
2. Ljungberg B, Campbell SC, Choi HY, et al. The epidemiology of renal cell carcinoma. *Eur Urol*. Oct 2011;60(4):615-621. [doi: [10.1016/j.eururo.2011.06.049](#)] [Medline: [21741761](#)]
3. Lipworth L, Tarone RE, Lund L, McLaughlin JK. Epidemiologic characteristics and risk factors for renal cell cancer. *Clin Epidemiol*. Aug 9, 2009;1:33-43. [doi: [10.2147/clep.s4759](#)] [Medline: [20865085](#)]
4. Chow WH, Dong LM, Devesa SS. Epidemiology and risk factors for kidney cancer. *Nat Rev Urol*. May 2010;7(5):245-257. [doi: [10.1038/nrurol.2010.46](#)] [Medline: [20448658](#)]
5. Haas NB, Nathanson KL. Hereditary kidney cancer syndromes. *Adv Chronic Kidney Dis*. Jan 2014;21(1):81-90. [doi: [10.1053/j.ackd.2013.10.001](#)] [Medline: [24359990](#)]
6. Hemminki K, Försti A, Hemminki A, Ljungberg B, Hemminki O. Progress in survival in renal cell carcinoma through 50 years evaluated in Finland and Sweden. *PLoS One*. 2021;16(6):e0253236. [doi: [10.1371/journal.pone.0253236](#)] [Medline: [34157049](#)]
7. Bukavina L, Bensalah K, Bray F, et al. Epidemiology of renal cell carcinoma: 2022 update. *Eur Urol*. Nov 2022;82(5):529-542. [doi: [10.1016/j.eururo.2022.08.019](#)] [Medline: [36100483](#)]
8. Aweys H, Lewis D, Sheriff M, et al. Renal cell cancer - insights in drug resistance mechanisms. *Anticancer Res*. Nov 2023;43(11):4781-4792. [doi: [10.21873/anticancerres.16675](#)] [Medline: [37909991](#)]
9. Cowey CL, Rathmell WK. VHL gene mutations in renal cell carcinoma: role as a biomarker of disease outcome and drug efficacy. *Curr Oncol Rep*. Mar 2009;11(2):94-101. [doi: [10.1007/s11912-009-0015-5](#)] [Medline: [19216840](#)]
10. Cairns P. Renal cell carcinoma. *Cancer Biomark*. 2011;9(1-6):461-473. [doi: [10.3233/CBM-2011-0176](#)]
11. Choueiri TK, Pal SK, Lewis B, Poteat S, Pels K, Hammers H. The 5th Kidney Cancer Research Summit: research accelerating cures for renal cell carcinoma in 2023. *Oncologist*. Feb 2, 2024;29(2):91-98. [doi: [10.1093/oncolo/oyad322](#)] [Medline: [38048064](#)]



12. Matorevhu A. Bibliometrics: application opportunities and limitations. In: *Bibliometrics - An Essential Methodological Tool for Research Projects*. IntechOpen; 2024. [doi: [10.5772/intechopen.1005292](https://doi.org/10.5772/intechopen.1005292)]
13. Escobar KM, Vicente-Villardón JL, de la Hoz-M J, Useche-Castro LM, Alarcón Cano DF, Siteneski A. Frequency of neuroendocrine tumor studies: using latent Dirichlet allocation and HJ-biplot statistical methods. *Mathematics*. 2021;9(18):2281. [doi: [10.3390/math9182281](https://doi.org/10.3390/math9182281)]
14. De La Hoz-M J, Mendes S, Fernández-Gómez MJ, González Silva Y. Capturing the complexity of COVID-19 research: trend analysis in the first two years of the pandemic using a Bayesian probabilistic model and machine learning tools. *Computation*. 2022;10(9):156. [doi: [10.3390/computation10090156](https://doi.org/10.3390/computation10090156)]
15. Pilacuan-Bonete L, Galindo-Villardón P, Delgado-Álvarez F. HJ-Biplot as a tool to give an extra analytical boost for the latent Dirichlet assignment (LDA) model: with an application to digital news analysis about COVID-19. *Mathematics*. 2022;10(14):2529. [doi: [10.3390/math10142529](https://doi.org/10.3390/math10142529)]
16. Yu D, Fang A, Xu Z. Topic research in fuzzy domain: Based on LDA topic modelling. *Inf Sci*. Nov 2023;648:119600. [doi: [10.1016/j.ins.2023.119600](https://doi.org/10.1016/j.ins.2023.119600)]
17. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. *J Mach Learn Res*. 2003;3:993-1022. URL: <https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf> [Accessed 2025-12-24]
18. Sood A, Ghosh AK. Literature search using PubMed: an essential tool for practicing evidence- based medicine. *J Assoc Physicians India*. Apr 2006;54:303-308. [Medline: [16944614](https://pubmed.ncbi.nlm.nih.gov/16944614/)]
19. Burnham JF. Scopus database: a review. *Biomed Digit Libr*. Mar 8, 2006;3(1):1. [doi: [10.1186/1742-5581-3-1](https://doi.org/10.1186/1742-5581-3-1)] [Medline: [16522216](https://pubmed.ncbi.nlm.nih.gov/16522216/)]
20. The R Project for Statistical Computing. URL: <https://www.r-project.org/> [Accessed 2025-12-24]
21. Grün B, Hornik K. Topicmodels: an R package for fitting topic models. *J Stat Softw*. 2011;40:1-30. [doi: [10.18637/jss.v040.i13](https://doi.org/10.18637/jss.v040.i13)]
22. Aria M, Cuccurullo C. bibliometrix: An R-tool for comprehensive science mapping analysis. *J Informetr*. Nov 2017;11(4):959-975. [doi: [10.1016/j.joi.2017.08.007](https://doi.org/10.1016/j.joi.2017.08.007)]
23. Caputo A, Kargina M. A user-friendly method to merge Scopus and Web of Science data during bibliometric analysis. *J Market Anal*. Mar 2022;10(1):82-88. [doi: [10.1057/s41270-021-00142-7](https://doi.org/10.1057/s41270-021-00142-7)]
24. Donthu N, Kumar S, Mukherjee D, Pandey N, Lim WM. How to conduct a bibliometric analysis: an overview and guidelines. *J Bus Res*. Sep 2021;133:285-296. [doi: [10.1016/j.jbusres.2021.04.070](https://doi.org/10.1016/j.jbusres.2021.04.070)]
25. Glendowne P, Glendowne D. Interpretability of API call topic models: an exploratory study. Presented at: 53rd Hawaii International Conference on System Sciences; Jan 7-10, 2020; Maui, HI. [doi: [10.24251/HICSS.2020.793](https://doi.org/10.24251/HICSS.2020.793)]
26. Syed S, Spruit M. Full-text or abstract? Examining topic coherence scores using latent Dirichlet allocation. Presented at: 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA); Oct 19-21, 2017:165-174; Tokyo, Japan. [doi: [10.1109/DSAA.2017.61](https://doi.org/10.1109/DSAA.2017.61)]
27. Tang J, Chang Y, Liu H. Mining social media with social theories. *SIGKDD Explor Newsl*. Jun 16, 2014;15(2):20-29. [doi: [10.1145/2641190.2641195](https://doi.org/10.1145/2641190.2641195)]
28. De la Hoz-M J, Fernández-Gómez MJ, Mendes S. LDAShiny: an R package for exploratory review of scientific literature based on a Bayesian probabilistic model and machine learning tools. *Mathematics*. 2021;9(14):1671. [doi: [10.3390/math9141671](https://doi.org/10.3390/math9141671)]
29. Blei DM, Lafferty JD. Dynamic topic models. Presented at: ICML '06: Proceedings of the 23rd International Conference on Machine Learning; Jun 25-29, 2006:113-120; Pittsburgh, PA. [doi: [10.1145/1143844.1143859](https://doi.org/10.1145/1143844.1143859)]
30. Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell*. Jun 1984;6(6):721-741. [doi: [10.1109/tpami.1984.4767596](https://doi.org/10.1109/tpami.1984.4767596)] [Medline: [22499653](https://pubmed.ncbi.nlm.nih.gov/22499653/)]
31. Röder M, Both A, Hinneburg A. Exploring the space of topic coherence measures. Presented at: WSDM '15: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining; Feb 2-6, 2015:399-408; Shanghai, China. [doi: [10.1145/2684822.2685324](https://doi.org/10.1145/2684822.2685324)]
32. Lau JH, Grieser K, Newman D, Baldwin T. Automatic labelling of topic models. Presented at: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies; Jun 19-24, 2011:1536-1545; Portland, OR. URL: <https://aclanthology.org/P11-1154/> [Accessed 2025-12-19]
33. Xiong H, Cheng Y, Zhao W, Liu J. Analyzing scientific research topics in manufacturing field using a topic model. *Comput Ind Eng*. Sep 2019;135:333-347. [doi: [10.1016/j.cie.2019.06.010](https://doi.org/10.1016/j.cie.2019.06.010)]
34. Griffiths TL, Steyvers M. Finding scientific topics. *Proc Natl Acad Sci USA*. Apr 6, 2004;101(Suppl 1):5228-5235. [doi: [10.1073/pnas.0307752101](https://doi.org/10.1073/pnas.0307752101)] [Medline: [14872004](https://pubmed.ncbi.nlm.nih.gov/14872004/)]
35. Kaufman L, Rousseeuw PJ. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons; 1990. [doi: [10.1002/9780470316801](https://doi.org/10.1002/9780470316801)] ISBN: 9780471878766

36. Quddus MB, Pratt N, Nabi G. Chromosomal aberrations in renal cell carcinoma: an overview with implications for clinical practice. *Urol Ann.* 2019;11(1):6-14. [doi: [10.4103/UA.UA\\_32\\_18](https://doi.org/10.4103/UA.UA_32_18)] [Medline: [30787564](https://pubmed.ncbi.nlm.nih.gov/30787564/)]
37. Testa U, Pelosi E, Castelli G. Genetic alterations in renal cancers: identification of the mechanisms underlying cancer initiation and progression and of therapeutic targets. *Medicines (Basel).* Jul 29, 2020;7(8):44. [doi: [10.3390/medicines7080044](https://doi.org/10.3390/medicines7080044)] [Medline: [32751108](https://pubmed.ncbi.nlm.nih.gov/32751108/)]
38. Li QK, Pavlovich CP, Zhang H, Kinsinger CR, Chan DW. Challenges and opportunities in the proteomic characterization of clear cell renal cell carcinoma (ccRCC): a critical step towards the personalized care of renal cancers. *Semin Cancer Biol.* Apr 2019;55:8-15. [doi: [10.1016/j.semcancer.2018.06.004](https://doi.org/10.1016/j.semcancer.2018.06.004)] [Medline: [30055950](https://pubmed.ncbi.nlm.nih.gov/30055950/)]
39. Sharma R, Kannourakis G, Prithviraj P, Ahmed N. Precision medicine: an optimal approach to patient care in renal cell carcinoma. *Front Med (Lausanne).* 2022;9:766869. [doi: [10.3389/fmed.2022.766869](https://doi.org/10.3389/fmed.2022.766869)] [Medline: [35775004](https://pubmed.ncbi.nlm.nih.gov/35775004/)]
40. Ghafouri-Fard S, Shirvani-Farsani Z, Branicki W, Taheri M. MicroRNA signature in renal cell carcinoma. *Front Oncol.* 2020;10:596359. [doi: [10.3389/fonc.2020.596359](https://doi.org/10.3389/fonc.2020.596359)] [Medline: [33330087](https://pubmed.ncbi.nlm.nih.gov/33330087/)]
41. Yang L, Zou X, Zou J, Zhang GA. A review of recent research on the role of microRNAs in renal cancer. *Med Sci Monit.* May 8, 2021;27:e930639. [doi: [10.12659/MSM.930639](https://doi.org/10.12659/MSM.930639)] [Medline: [33963171](https://pubmed.ncbi.nlm.nih.gov/33963171/)]
42. Muglia VF, Prando A. Renal cell carcinoma: histological classification and correlation with imaging findings. *Radiol Bras.* 2015;48(3):166-174. [doi: [10.1590/0100-3984.2013.1927](https://doi.org/10.1590/0100-3984.2013.1927)] [Medline: [26185343](https://pubmed.ncbi.nlm.nih.gov/26185343/)]
43. Athanazio DA, Amorim LS, da Cunha IW, et al. Classification of renal cell tumors – current concepts and use of ancillary tests: recommendations of the Brazilian Society of Pathology. *Surg Exp Pathol.* Dec 2021;4(1):4. [doi: [10.1186/s42047-020-00084-x](https://doi.org/10.1186/s42047-020-00084-x)]
44. Sanguedolce F, Mazzucchelli R, Falagario UG, et al. Diagnostic biomarkers in renal cell tumors according to the latest WHO classification: a focus on selected new entities. *Cancers (Basel).* May 13, 2024;16(10):1856. [doi: [10.3390/cancers16101856](https://doi.org/10.3390/cancers16101856)] [Medline: [38791935](https://pubmed.ncbi.nlm.nih.gov/38791935/)]
45. Lam JS, Shvarts O, Leppert JT, Pantuck AJ, Figlin RA, Belldgrun AS. Postoperative surveillance protocol for patients with localized and locally advanced renal cell carcinoma based on a validated prognostic nomogram and risk group stratification system. *J Urol.* Aug 2005;174(2):466-472. [doi: [10.1097/01.ju.0000165572.38887.da](https://doi.org/10.1097/01.ju.0000165572.38887.da)] [Medline: [16006866](https://pubmed.ncbi.nlm.nih.gov/16006866/)]
46. Potti A, George DJ. Tyrosine kinase inhibitors in renal cell carcinoma. *Clin Cancer Res.* Sep 15, 2004;10(18 Pt 2):6371S-6S. [doi: [10.1158/1078-0432.CCR-050014](https://doi.org/10.1158/1078-0432.CCR-050014)] [Medline: [15448033](https://pubmed.ncbi.nlm.nih.gov/15448033/)]
47. Schöffski P, Dumez H, Clement P, et al. Emerging role of tyrosine kinase inhibitors in the treatment of advanced renal cell cancer: a review. *Ann Oncol.* Aug 2006;17(8):1185-1196. [doi: [10.1093/annonc/mdj133](https://doi.org/10.1093/annonc/mdj133)] [Medline: [16418310](https://pubmed.ncbi.nlm.nih.gov/16418310/)]
48. Xu W, Atkins MB, McDermott DF. Checkpoint inhibitor immunotherapy in kidney cancer. *Nat Rev Urol.* Mar 2020;17(3):137-150. [doi: [10.1038/s41585-020-0282-3](https://doi.org/10.1038/s41585-020-0282-3)] [Medline: [32020040](https://pubmed.ncbi.nlm.nih.gov/32020040/)]
49. Rassy E, Flippot R, Albiges L. Tyrosine kinase inhibitors and immunotherapy combinations in renal cell carcinoma. *Ther Adv Med Oncol.* 2020;12:1758835920907504. [doi: [10.1177/1758835920907504](https://doi.org/10.1177/1758835920907504)] [Medline: [32215057](https://pubmed.ncbi.nlm.nih.gov/32215057/)]
50. Braun DA, Bakouny Z, Hirsch L, et al. Beyond conventional immune-checkpoint inhibition - novel immunotherapies for renal cell carcinoma. *Nat Rev Clin Oncol.* Apr 2021;18(4):199-214. [doi: [10.1038/s41571-020-00455-z](https://doi.org/10.1038/s41571-020-00455-z)] [Medline: [33437048](https://pubmed.ncbi.nlm.nih.gov/33437048/)]
51. Ravaud A. Treatment-associated adverse event management in the advanced renal cell carcinoma patient treated with targeted therapies. *Oncologist.* 2011;16 Suppl 2(Suppl 2):32-44. [doi: [10.1634/theoncologist.2011-S2-32](https://doi.org/10.1634/theoncologist.2011-S2-32)] [Medline: [21346038](https://pubmed.ncbi.nlm.nih.gov/21346038/)]
52. Kamli H, Li L, Gobe GC. Limitations to the therapeutic potential of tyrosine kinase inhibitors and alternative therapies for kidney cancer. *Ochsner J.* 2019;19(2):138-151. [doi: [10.31486/toj.18.0015](https://doi.org/10.31486/toj.18.0015)] [Medline: [31258426](https://pubmed.ncbi.nlm.nih.gov/31258426/)]
53. Espi M, Teuma C, Novel-Catin E, et al. Renal adverse effects of immune checkpoints inhibitors in clinical practice: ImmuNoTox study. *Eur J Cancer.* Apr 2021;147:29-39. [doi: [10.1016/j.ejca.2021.01.005](https://doi.org/10.1016/j.ejca.2021.01.005)] [Medline: [33607383](https://pubmed.ncbi.nlm.nih.gov/33607383/)]
54. Barrueto L, Caminero F, Cash L, Makris C, Lamichhane P, Deshmukh RR. Resistance to checkpoint inhibition in cancer immunotherapy. *Transl Oncol.* Mar 2020;13(3):100738. [doi: [10.1016/j.tranon.2019.12.010](https://doi.org/10.1016/j.tranon.2019.12.010)] [Medline: [32114384](https://pubmed.ncbi.nlm.nih.gov/32114384/)]
55. Sato K, Tsuchiya N, Sasaki R, et al. Increased serum levels of vascular endothelial growth factor in patients with renal cell carcinoma. *Jpn J Cancer Res.* Aug 1999;90(8):874-879. [doi: [10.1111/j.1349-7006.1999.tb00829.x](https://doi.org/10.1111/j.1349-7006.1999.tb00829.x)] [Medline: [10543260](https://pubmed.ncbi.nlm.nih.gov/10543260/)]
56. Schips L, Dalpiaz O, Lipsky K, et al. Serum levels of vascular endothelial growth factor (VEGF) and endostatin in renal cell carcinoma patients compared to a control group. *Eur Urol.* Jan 2007;51(1):168-173. [doi: [10.1016/j.eururo.2006.06.026](https://doi.org/10.1016/j.eururo.2006.06.026)] [Medline: [16844285](https://pubmed.ncbi.nlm.nih.gov/16844285/)]
57. Beuselinck B, Vano YA, Oudard S, et al. Prognostic impact of baseline serum C-reactive protein in patients with metastatic renal cell carcinoma (RCC) treated with sunitinib. *BJU Int.* Jul 2014;114(1):81-89. [doi: [10.1111/bju.12494](https://doi.org/10.1111/bju.12494)] [Medline: [24215209](https://pubmed.ncbi.nlm.nih.gov/24215209/)]

58. Pantuck AJ, Zeng G, Belldegrun AS, Figlin RA. Pathobiology, prognosis, and targeted therapy for renal cell carcinoma: exploiting the hypoxia-induced pathway. Clin Cancer Res. Oct 15, 2003;9(13):4641-4652. [Medline: [14581333](#)]
59. Yu L, Qiu Y, Tong X. Ferroptosis in renal cancer therapy: a narrative review of drug candidates. Cancers (Basel). Sep 11, 2024;16(18):3131. [doi: [10.3390/cancers16183131](#)] [Medline: [39335103](#)]
60. Kruk L, Mamtimin M, Braun A, et al. Inflammatory networks in renal cell carcinoma. Cancers (Basel). Apr 9, 2023;15(8):2212. [doi: [10.3390/cancers15082212](#)] [Medline: [37190141](#)]
61. Yang JW, Wan S, Li KP, Chen SY, Yang L. Gut and urinary microbiota: the causes and potential treatment measures of renal cell carcinoma. Front Immunol. 2023;14:1188520. [doi: [10.3389/fimmu.2023.1188520](#)] [Medline: [37441065](#)]
62. Sun M, Shariat SF, Cheng C, et al. Prognostic factors and predictive models in renal cell carcinoma: a contemporary review. Eur Urol. Oct 2011;60(4):644-661. [doi: [10.1016/j.eururo.2011.06.041](#)] [Medline: [21741163](#)]
63. Hsieh JJ, Le V, Cao D, Cheng EH, Creighton CJ. Genomic classifications of renal cell carcinoma: a critical step towards the future application of personalized kidney cancer care with pan-omics precision. J Pathol. Apr 2018;244(5):525-537. [doi: [10.1002/path.5022](#)] [Medline: [29266437](#)]
64. Knudsen JE, Rich JM, Ma R. Artificial intelligence in pathomics and genomics of renal cell carcinoma. Urol Clin North Am. Feb 2024;51(1):47-62. [doi: [10.1016/j.ucl.2023.06.002](#)] [Medline: [37945102](#)]
65. Papadia G, Pacella M, Perrone M, Giliberti V. A comparison of different topic modeling methods through a real case study of Italian customer care. Algorithms. 2023;16(2):94. [doi: [10.3390/a16020094](#)]

## Abbreviations

**LDA:** latent Dirichlet allocation  
**MeSH:** Medical Subject Headings  
**RCC:** renal cell carcinoma  
**TKI:** tyrosine kinase inhibitor

*Edited by Naomi Cahill; peer-reviewed by G Fan, L Raymond Guo, Nicolas Bievre; submitted 10.Jun.2025; final revised version received 31.Oct.2025; accepted 31.Oct.2025; published 16.Jan.2026*

### Please cite as:

De La Hoz-M J, Montes-Escobar K, Salas-Macias CA, Fors M, Ballaz SJ

Using Latent Dirichlet Allocation Topic Modeling to Uncover Latent Research Topics and Trends in Renal Cell Carcinoma: Bibliometric Review

JMIR Cancer 2026;12:e78797

URL: <https://cancer.jmir.org/2026/1/e78797>

doi: [10.2196/78797](#)

© Javier De La Hoz-M, Karime Montes-Escobar, Carlos Alfredo Salas-Macias, Martha Fors, Santiago J Ballaz. Originally published in JMIR Cancer (<https://cancer.jmir.org>), 16.Jan.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Cancer, is properly cited. The complete bibliographic information, a link to the original publication on <https://cancer.jmir.org/>, as well as this copyright and license information must be included.