Heidelberg Hospital University

Machine-learning-based bibliometric analysis of pancreatic cancer research over the past 25 years

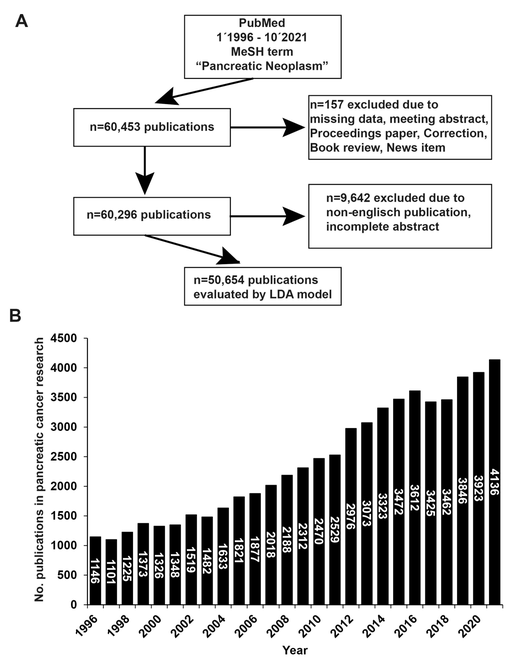
Kangtao Wang, Ingrid Herr

Molecular OncoSurgery, Section Surgical Research, Department of General, Visceral and Transplantation Surgery, University of Heidelberg, Heidelberg, Germany

Abstract

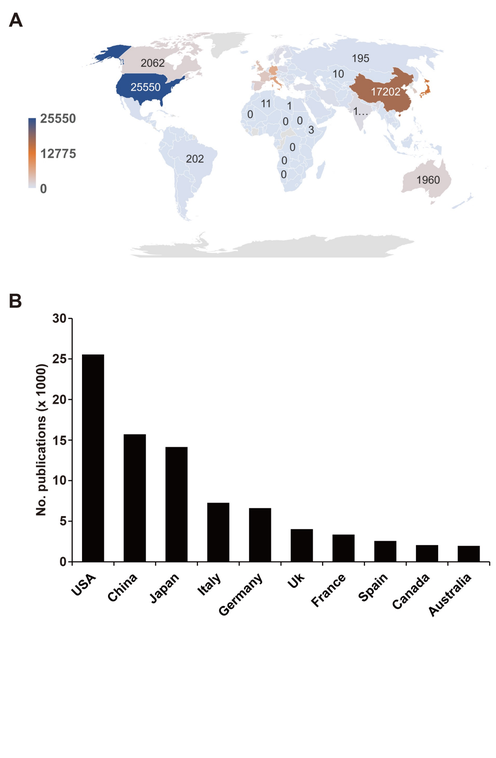
Machine learning and semantic analysis are computer-based methods to evaluate complex relationships and predict future perspectives. We used these technologies to define recent, current and future topics in pancreatic cancer research. Publications indexed under the Medical Subject Headings (MeSH) term 'Pancreatic Neoplasms' from January 1996 to October 2021 were downloaded from PubMed. Using the statistical computing language R and the interpreted, high-level, general-purpose programming language Python, we extracted publication dates, geographic information, and abstracts from each publication's metadata for bibliometric analyses. The generative statistical algorithm "latent Dirichlet allocation" (LDA) was applied to identify specific research topics and trends. The unsupervised "Louvain algorithm" was used to establish a network to identify relationships between single topics. A total of 60,296 publications were identified and analyzed. The publications were derived from 133 countries, mostly from the Northern Hemisphere. For the term "pancreatic cancer research", 12,058 MeSH terms appeared 1,395,060 times. Among them, we identified the four main topics "Clinical Manifestation and Diagnosis", "Review and Management", "Treatment Studies", and "Basic Research". The number of publications has increased rapidly during the past 25 years. Based on the number of publications, the algorithm predicted that "Immunotherapy", Prognostic research", "Protein expression", "Case reports", "Gemcitabine and mechanism", "Clinical study of gemcitabine", "Operation and postoperation", "Chemotherapy and resection", and "Review and management" as current research topics. To our knowledge, this is the first study on this subject of pancreatic cancer research, which has become possible due to the improvement of algorithms and hardware.

1 The number of publications on pancreatic cancer has increased rapidly in the past 25 years.



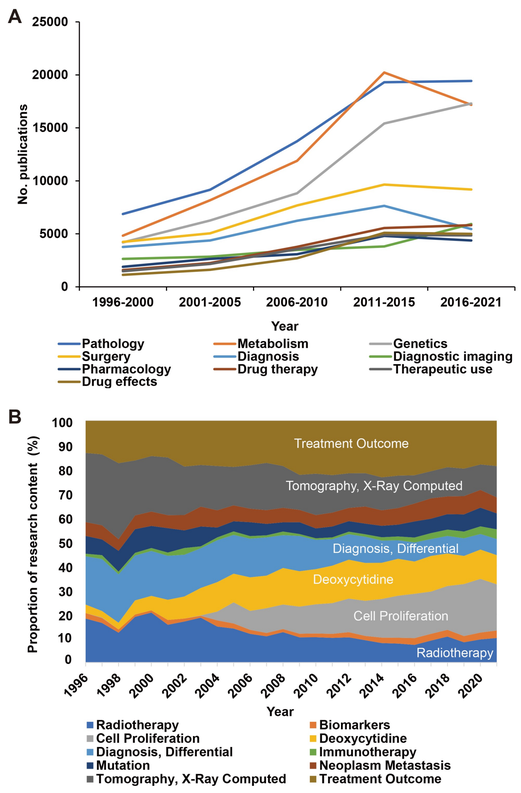
(A) Figure 1. The number of publications on pancreatic cancer has increased rapidly in the past 25 years. (A) PubMed publications from January 1st, 1996 to October 10th, 2021 were screened and downloaded using the R package easyPubMed and the MeSH term "Pancreatic Neoplasms". From the initially identified 60.453 publications, 157 publications were excluded manually due to missing data or when the publication was a meeting abstract, proceedings paper, a correction, a book review, or a news item. Of the resulting 60,296 publications, another 9,642 were manually excluded when the language was not English or the abstract was incomplete, resulting in 50,654 publications. (B) The selected 50,654 publications were analyzed by LDA and Python. The data were visualized with Excel and R. The number of publications (No. publications) per year is shown.

2 Global pancreatic cancer research differs significantly between regions



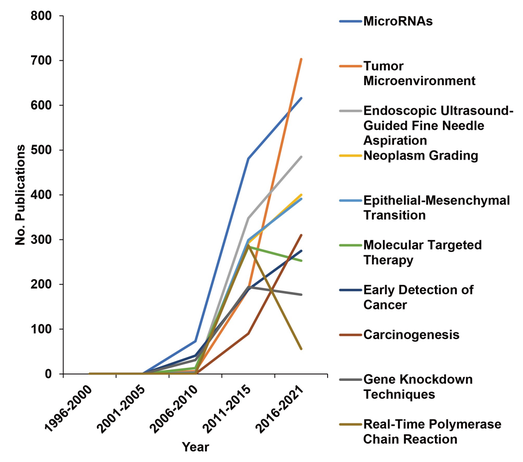
(A) The global distribution of pancreatic cancer publications in the last 25 years by number is shown. We extracted country information based on the publication’s affiliation. The darker the color is, the greater the volume of publications. Significant head effect: The number of posts in the Northern Hemisphere is much higher than that in the Southern Hemisphere. **(B)** Top 10 countries with the highest publication numbers in pancreatic cancer research.

3 The MeSH terms "Pathology" and "Metabolism" occur most frequently



(A) R extracted significant MeSH terms from publications listed in the PubMed database according to the publication topic. (A) Ten of the most widely studied MeSH terms and their number (No.) of publications per year. "Pathology", "metabolism", and "surgery" were the most frequently studied research fields in the last 25 years. (B) Proportional changes in some representative MeSH terms are given in percent (%) per year. The MeSH terms are marked by different colors as indicated.

4 Changes in the 10 most widely studied newly merged MeSH terms over the years



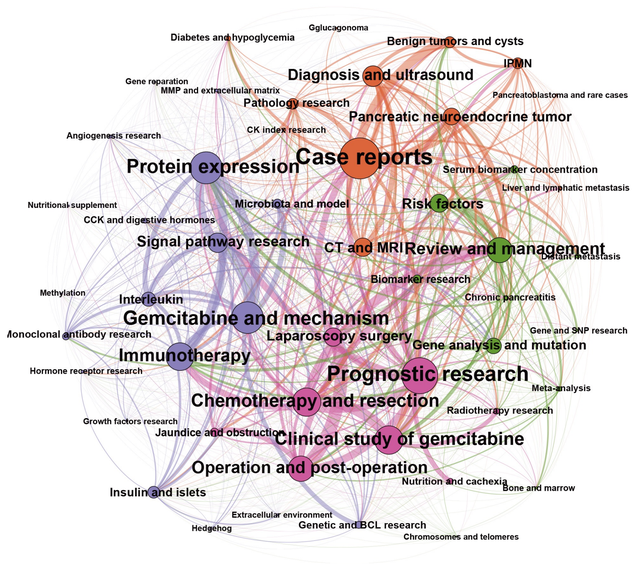
MeSH terms were extracted by R, and newly merged MeSH terms that appeared after 1996 are shown as the number (No.) of publications per year.

Novelty & Impact

The number of publications in pancreatic cancer research has increased rapidly during the past 25 years, and it is a great challenge for scientists to filter out important information from the exploding number of publications. Using machine learning and natural language processing, we identified the research topics that developed from "Interleukin and Pathology Research" in the past, to "Immunotherapy", "Prognostics", "Microbiome", "Early Diagnosis", and "Molecular Typing" in the present and future. From our data we conclude that the future breakthrough in pancreatic cancer research may depend on the understanding of complex relationships, new technologies, and the application and popularization of new and innovative diagnostic technologies. We found only a few studies with the topics "Hospice Care", "Quality of Life", "Patient Perspectives" and "Economics", suggesting that these fields are not currently being intensively researched. Together, machine learning based on perfect medical record text, publication databases, and improved algorithms, may be reasonable to play an active role in future literature search and evaluation of research directions. Machine learning and Natural Language Processing may be a handy new tool for scientists, to extract objective and comprehensive clues from huge data amounts.

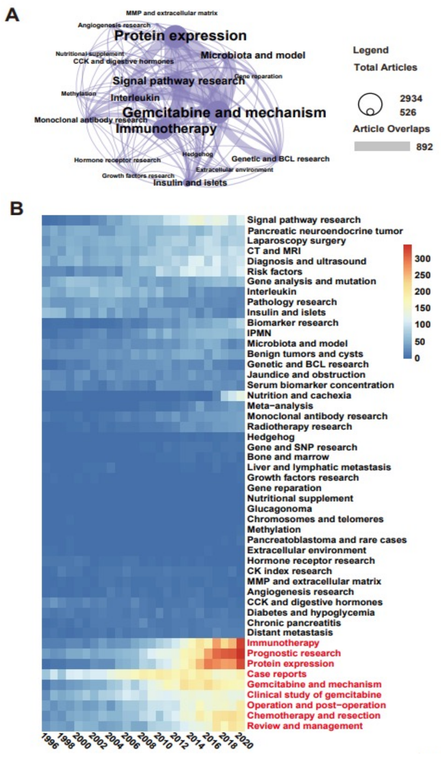
Contact: i.herr@uni-heidelberg.de

5 The main areas in pancreatic cancer research over the last 25 years are “Treatment Studies”, “Clinical Manifestations and Diagnosis”, “Review and Management”, and "Basic Research".



(A) MeSH terms were extracted by R, and newly merged MeSH terms that appeared after 1996 are shown as the number (No.) of We analyzed pancreatic cancer publications from the last 25 years using the LDA algorithm, from which 50 topics were identified and classified into four broad categories: "Clinical Manifestations and Diagnosis" (orange), "Review and Management" (green), "Treatment Study" (pink), and "Basic Research" (purple). The circle size represents the number of publications contained in the research topic. For example, the algorithm found that the topic case reports contain 3,791 publications. The line thickness between the circles represents the degree of overlap between the two research topics, such as the 989 publications with the highest degree of overlap in "Operation and postoperation" and "Laparoscopy surgery".

6 "Immunotherapy" and "Prognostic research" are current research topics.



(A) Topic cluster network of the topic "Basic research" and the interrelation of subtopics was analyzed using the LDA algorithm. In this cluster, the "gemcitabine and mechanism" topic included 2,934 publications, and 892 of these publications shared the same topic with the topic "protein expression". **(B)** The heatmap presents the number of publications per year of 50 research topics over the past 25 years. The data were generated by the use of the LDA algorithm. The abscissa represents the year, the ordinate represents the topics, and the color brightness represents the number of publications and reflects the shift in research focus.