Research trends analysis of chronic hepatitis C versus nonalcoholic fatty liver disease: A literature review text-mining analysis of publications

Eyal Klang1,2,3 | Shelly Soffer3,4,5 | Lee Alper3,6 | Orit Shimon7 | Yiftach Barash1,3 | Yana Davidov8 | Mariya Likhter8 | Oranit Cohen-Ezra8 | Gil Ben Yakov8 | Ziv Ben-Ari6,8

Abstract

Background: Hepatitis C virus (HCV) rates have lowered due to direct-acting antiviral treatment. Nonalcoholic steatohepatitis (NASH)/nonalcoholic fatty liver disease (NAFLD) is rising with no available therapy. We employed text-mining to analyze trends in HCV and NAFLD research from the past two decades.

Materials and Methods: We queried PubMed for all HCV and NASH/NAFLD entries published between 2000 and 2020. We compared the total number of publications on both etiologies. We performed subanalyses for different terms of interest and for geographic origin.

Results: Overall, 75,934 HCV-related entries and 24,987 NASH/NAFLD-related entries were published during the study period. Up to 2015, there was a linear upward slope in the number of annual HCV publications (154.9 publications/year, \( p < 0.001 \)). In 2015, the yearly number of HCV publications started showing a downward slope (−242.2 publications/year, \( p < 0.001 \)). The number of NASH/NAFLD publications showed a continuous upward slope during the study period. The NASH/NAFLD field lacks publications on screening and treatment methods.

Conclusion: Trends in publications varied between both etiologies. They reflect the success of antiviral treatment for HCV. The growing rates of NAFLD/NASH and the lack of a targeted cure explain the rise in related publications.

Keywords: chronic hepatitis C, nonalcoholic fatty liver disease, PubMed, text mining
**INTRODUCTION**

Nonalcoholic fatty liver disease (NAFLD) is rapidly becoming the most common cause of liver disease in Western populations. A recent meta-analysis estimated the pooled prevalence of NAFLD is 33.8%. The spectrum of nonalcoholic liver diseases includes nonalcoholic fatty liver (simple steatosis) and its progressive form, nonalcoholic steatohepatitis (NASH), differentiated by hepatic inflammation and fibrosis in the latter. Patients can progress to liver cirrhosis and primary liver cancer. At present, the most effective first-line therapy for the management of NAFLD/NASH is lifestyle modifications with diet and exercise. However, long-term adherence to lifestyle modification is rare in the target population. Other available therapies include Metformin and vitamin E, specifically for NASH patients with at least F3 fibrosis. For NASH patients with diabetes, possible treatments are GLP-1 agonists and pioglitazone. Currently, a multitude of novel treatments is under development for wider patient populations.

Hepatitis B and Hepatitis C are the major etiologies of cirrhosis and hepatocellular carcinoma (HCC). Although the prevalence of hepatitis C virus (HCV) is declining, chronic hepatitis C (CHC) is still one of the major etiologies of cirrhosis and HCC in the world. It has been estimated that approximately 58 million individuals worldwide have CHC. The reduction in CHC rates is due to the highly effective direct-acting antivirals (DAAs), first introduced in 2013, which have been associated with sustained virological response in more than 95% of treated patients. The World Health Organization (WHO) set an ambitious goal in 2016 to eliminate hepatitis C as a major public health threat by 2030. The total number of publications on the above-mentioned topics in the field of hepatology has markedly increased over the past decades, rendering it impossible to manually summarize each topic or compare research trends.

Text-mining is a computer-based technique for deriving information from free text. We employed text-mining techniques to delineate the main trends in published HCV and NAFLD research, to understand publication gaps and to point out the need for potential future investigations.

**MATERIALS AND METHODS**

**2.1 Data set**

The US National Center for Biotechnology Information (NCBI) provides public application programming interfaces (APIs) that enable programmatic access to the PubMed database. This study used the publicly available PyMed Python package to query the PubMed API.

The following data were extracted for each entry: PubMed unique article ID (PMID), title, publishing journal, abstract text, keywords (if any), and authors’ affiliations. Data lock was performed on April 25, 2021.

**2.2 Inclusion criteria**

The entire MEDLINE/PubMed database was used as the source for this analysis. Two searches of entry titles and abstracts were performed. The first search was for the terms “hepatitis c” OR “HCV.” The second search was for the terms “nonalcoholic steatohepatitis” OR “nonalcoholic fatty liver disease” OR “nonalcoholic steatohepatitis” OR “nonalcoholic fatty liver disease.” These terms were matched to the tokenized title list and a subset of records was retrieved.

To analyze the annual publication rates in the past two decades, the entries were limited to those published between January 1, 2000, and December 31, 2020.

**2.3 Data processing**

The data processing and result visualization were written in Python (Ver. 3.6.5, 64 bits). For text-mining, each title, study abstract, and authors’ affiliations were lowercased.

**TABLE 1 Total number of publications (2000–2020) by different terms of interest in the NASH/NAFLD group**

| Terms NAFLD | Publications (N) | Rate (%) |
|-------------|-----------------|----------|
| Fibrosis, cirrhosis | 8666 | 34.7 |
| Obesity | 6273 | 25.1 |
| Metabolic syndrome | 3740 | 15.0 |
| Prevalence | 3671 | 14.7 |
| Diagnosis | 2549 | 10.2 |
| Risk factors | 2321 | 9.3 |
| Noninvasive, biomarkers, elastography | 2149 | 8.6 |
| Liver biopsy | 2145 | 8.6 |
| Diabetes mellitus | 1839 | 7.4 |
| Hypertension | 1359 | 5.4 |
| Imaging | 1142 | 4.6 |
| Dyslipidemia | 1091 | 4.4 |
| Screening | 924 | 3.7 |
| Bariatric | 653 | 2.6 |
| Prognosis | 507 | 2.0 |
| Lean | 444 | 1.8 |
| Epidemiology | 378 | 1.5 |
| Surveillance | 244 | 1.0 |
| Genetics | 193 | <1 |
| Pharmacotherapy | 150 | <1 |
| Lifestyle | 53 | <1 |

Abbreviations: NAFLD, nonalcoholic fatty liver disease; NASH, nonalcoholic steatohepatitis.
First, the total number of publications in the HCV and the NASH/NAFLD groups was compared. Then, a subanalysis was performed by searching each entry title and abstract for the terms: “treatment,” “randomized controlled trial/RCT,” “meta-analysis/metaanalysis,” “multi-center/multicenter,” and “screening.” Subanalyses for other terms of interest are presented in Tables 1 and 2.

A subanalysis by country of origin was also performed. All countries were retrieved from the affiliation data. The total numbers of publications by country were evaluated separately for HCV and NASH/NAFLD publications, with some entries affiliated with more than one country. The total number of publications was also normalized by country population in 2020, as extracted from the World Bank Data Catalog (https://datacatalog.worldbank.org/).

2.4 Statistical analysis

All analyses were conducted with Python (Python software foundation, Version 3.6.5). Statistical significance was established at a two-sided $p < 0.05$. Descriptive statistics were reported using counts with percentages for categorical variables.

Annual trends of publications for the years 2000–2020 were plotted for the different analyses. The slopes of publication trends were calculated by fitting linear regression lines to the annual number of publications in the years 2000–2020 (with $X$ being the calendar year and $Y$ being the annual publications count). $p$ values were calculated for the linear regression lines.

### RESULTS

Out of 31,850,051 PubMed records available on the search date, 75,934 (0.2%) HCV-related entries and 24,987 (0.08%) NASH/NAFLD-related entries were published during 2000–2020 (Figure 1). Of these, 1462 entries concerned both HCV and NASH/NAFLD.

For geographical analyses, 93.3% entries had a country affiliation.

#### 3.1 Publication trends

A comparison of trends of publications during 2000–2020, relating to HCV versus NASH/NAFLD, is presented in Figure 2. Up until approximately 2015, there was a linear upward slope of annual publications relating to HCV (154.9 publications/year, $p < 0.001$). This trend was reversed in 2015, when the annual number of HCV publications showed a downward slope (~242.2 publications/year, $p < 0.001$).

#### TABLE 2 Total number of publications (2000–2020) by different terms of interest in the HCV group

| Terms                                      | Publications (N) | Rate (%) |
|--------------------------------------------|------------------|----------|
| Fibrosis, cirrhosis                        | 16,984           | 22.4     |
| Hepatocellular carcinoma, HCC              | 9535             | 12.6     |
| Efficacy                                   | 5981             | 7.9      |
| Safety                                     | 4148             | 5.5      |
| Resistance                                 | 3875             | 5.1      |
| Direct acting antiviral agent, DAA         | 3579             | 4.7      |
| Cure                                       | 2410             | 3.2      |
| Sofosbuvir, velpatasvir                    | 2408             | 3.2      |
| Drug users, opioids                        | 2361             | 3.1      |
| Surveillance                               | 1580             | 2.1      |
| Decompensation                             | 734              | 1.0      |
| Drug–drug interaction                      | 468              | <1       |
| Inmates                                    | 272              | <1       |
| Glecaprevir, pibrentasvir                  | 248              | <1       |
| Direct acting antiviral agent failure, DAA failure | 45              | <1       |

Abbreviation: HCV, hepatitis C virus.

#### FIGURE 1 Study inclusion chart

#### FIGURE 2 A comparison of the overall HCV and NASH/NAFLD publication trends, displayed by the number of publications per year during 2000–2020. HCV, hepatitis C viral; NAFLD, nonalcoholic fatty liver disease; NASH, nonalcoholic steatohepatitis.
The annual number of publications relating to NASH/NAFLD continuously rose between 2000 and 2020. The magnitude of the slope increased between 2000 and 2015 (114.2 publications/year, \( p < 0.001 \)) and between 2015 and 2020 (439.7 publications/year, \( p = 0.01 \)). In 2020, the number of NASH/NAFLD publications surpassed the number of HCV publications.

The number of HCV and NASH/NAFLD publications with the term “treatment” followed these same trends, yet the number of NASH/NAFLD publications during 2020 was lower than those on HCV (Figure 3A).

Among the publications relating to “randomized controlled trials” and “meta-analysis,” NASH/NAFLD publications surpassed HCV publications (RCT in 2015, meta-analysis in 2019) and continued to increase exponentially (Figure 3B>C).

For both “multicenter” and “screening” search terms, the number of NASH/NAFLD-related publications showed a milder upward slope, with a total number of publications markedly lower than the number of multicenter/screening HCV studies (Figure 3D,E).

Analyses of other terms of interest for HCV and NASH/NAFLD are presented in Figure 4A,B.

**FIGURE 3** Annual HCV and NASH/NAFLD (2000–2020) publication trends for different terms of interest, displayed by the number of publications per year: treatment (A), randomized controlled trials (B), meta-analysis (C), multicenter (D), and screening (E). HCV, hepatitis C viral; NAFLD, nonalcoholic fatty liver disease; NASH, nonalcoholic steatohepatitis.
3.2 | Geographical analysis

The largest number of HCV-related publications were affiliated with research teams in the United States, with trends matching those of the overall publication trends (Figure 5A). Other countries with high numbers of HCV publications included Japan, Italy, China, and France. Both the United States and China were leading in the number of NASH/NAFLD-related publications, with an exponential curve over the study period. In 2020, the number of publications affiliated with institutions in China was equal to the number published by groups in the United States (Figure 5B). The numbers of publications per country normalized by population size are presented in Supporting Information: Tables 1 and 2.

4 | DISCUSSION

This study applied a text-mining approach to capture a snapshot of the trends in HCV and NAFLD/NASH research-related publications over the last 20 years. Trends in research publications were significantly different between these two major etiologies of chronic liver diseases. The decrease in the annual number of HCV publications from 2015 and on reflected the 2013 introduction of the highly successful DAAs.10

The number of NASH/NAFLD-related publications continuously rose during the study period, often parallel to the NASH/NAFLD prevalence and the awareness in countries where the epidemiological trends could be documented.13 The number of annual publications relating to NAFLD/NASH therapeutics is still increasing and a multitude of pharmacological targets are currently in development. The drug that is the most advanced in terms of clinical trial research is the farnesoid X receptor agonist, obeticholic acid.7 Despite such developments, identifying the most effective therapeutic intervention is challenging.

The number of "randomized controlled trials" and "meta-analysis" NASH/NAFLD publications surpassed HCV publications, also likely due to the rise in NASH/NAFLD prevalence. The number of publications on multicenter HCV-related studies was markedly higher than in similar NAFLD/NASH studies, possibly due to the introduction of effective treatment. In parallel, and likely for the same reasons, a marked increase in multicenter observational studies analyzing the DAAs in real-world settings14,15 were published. The higher number of screening publications for HCV may be attributed to the availability of a cure for hepatitis.
C, as opposed to NAFLD/NASH. Diagnosis of hepatitis C currently remains one of the major gaps in the cascade of care. Although several successful elimination efforts have been reported, research studies regarding effective screening strategies are required. One possible marker for the diagnosis of hepatitis is serum procalcitonin, which was found to be higher in patients with HCV-related cirrhosis. Since 2015, the number of NAFLD/NASH screening publications is also increasing. This trend will likely influence screening efforts on a global level and research trends.

As expected, the number of HCV publications relating to specific terms (fibrosis and cirrhosis, HCC, efficacy, safety, and resistance) peaked in 2016 and then sloped down. In contrast, the number of NAFLD/NASH publications relating to select terms increased exponentially between 2012 and 2020, with fibrosis, cirrhosis (34.7%), obesity (25.1%), metabolic syndrome (15%), and prevalence (14.7%) peaking mostly, all of which impact disease prognosis. Of note, these trends paralleled the recent increase in the global awareness and prevalence of the disease.

With regard to country-specific publication trends, the curve of HCV-related, United States-originating publications followed the overall publication trend.

Both the United States- and China-based teams generated the highest number of NASH/NAFLD publications, with trends exhibiting an exponential curve, likely due to the lack of effective treatment. The United States led in the number of HCV and NAFLD/NASH publications, reflecting the prominent role of the United States Hepatologists community in global hepatology research. The prevalence of NAFLD/NASH in China has been increasing over the years and is now as high as in the Western world. In fact, China is projected to experience the largest increase in the number of NAFLD cases in the coming years.

The presented analysis has several limitations. The sheer number of publications precludes a manual analysis of the records. Thus, this analysis only provided a high-level summary of publications. The list of search terms was determined based on current data in the study field. However, other terms might have reported on entirely different explanatory variables that correlate with national publication rates and might have yielded different results.

CONFLICT OF INTEREST
The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT
The data that support the findings of this study are openly available in PubMed at https://pubmed.ncbi.nlm.nih.gov/ using the DOIs or PMID listed as references. Eyal Klang had full access to all of the data in this study and takes complete responsibility for the integrity of the data and the accuracy of the data analysis.

TRANSPARENCY STATEMENT
Eyal Klang affirms that this manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

ORCID
Lee Alper http://orcid.org/0000-0001-8194-8300

REFERENCES
1. Younossi ZM, Blissett D, Blissett R, et al. The economic and clinical burden of nonalcoholic fatty liver disease in the United States and Europe. Hepatology. 2016;64(5):1577-1586. doi:10.1002/hep.28785
2. Goldberg D, Ditah IC, Saeian K, et al. Changes in the prevalence of hepatitis C virus infection, nonalcoholic steatohepatitis, and alcoholic liver disease among patients with cirrhosis or liver failure on the waitlist for liver transplantation. Gastroenterology. 2017;152(5):1090-1099. doi:10.1053/j.gastro.2017.01.003
3. Powell EE, Wong VW, Rinella M. Non-alcoholic fatty liver disease. Lancet. 2021;397(10290):2212-2224. doi:10.1016/S0140-6736(20)32511-3
4. Lim GEH, Tang A, Ng CH, et al. An observational data meta-analysis on the differences in prevalence and risk factors between NAFLD vs NAFLD. Clin Gastroenterol Hepatol. 2021;19(6):754-763. doi:10.1016/j.cgh.2021.01.038
5. Chalasani N, Younossi Z, Lavine JE, et al. The diagnosis and management of nonalcoholic fatty liver disease: practice guidance from the American Association for the Study of Liver Diseases. Hepatology. 2018;67(1):328-357. doi:10.1002/hep.29367

AUTHOR CONTRIBUTIONS
Eyal Klang: Conceptualization; data curation; formal analysis; investigation; supervision; writing—original draft; writing—review and editing. Shelly Soffer: Data curation; investigation; writing—review and editing. Lee Alper: Data curation; investigation; writing—review and editing. Orly Shimon: Data curation; investigation; writing—review and editing. Yiftach Barash: Data curation; investigation; writing—review and editing. Yana Davidov: Data curation; investigation; writing—review and editing. Mariya Likhter: Data curation; investigation; writing—review and editing. Orit Shimon: Data curation; investigation; writing—review and editing.

ORCID
Lee Alper http://orcid.org/0000-0001-8194-8300

REFERENCES
1. Younossi ZM, Blissett D, Blissett R, et al. The economic and clinical burden of nonalcoholic fatty liver disease in the United States and Europe. Hepatology. 2016;64(5):1577-1586. doi:10.1002/hep.28785
2. Goldberg D, Ditah IC, Saeian K, et al. Changes in the prevalence of hepatitis C virus infection, nonalcoholic steatohepatitis, and alcoholic liver disease among patients with cirrhosis or liver failure on the waitlist for liver transplantation. Gastroenterology. 2017;152(5):1090-1099. doi:10.1053/j.gastro.2017.01.003
3. Powell EE, Wong VW, Rinella M. Non-alcoholic fatty liver disease. Lancet. 2021;397(10290):2212-2224. doi:10.1016/S0140-6736(20)32511-3
4. Lim GEH, Tang A, Ng CH, et al. An observational data meta-analysis on the differences in prevalence and risk factors between NAFLD vs NAFLD. Clin Gastroenterol Hepatol. 2021;19(6):754-763. doi:10.1016/j.cgh.2021.01.038
5. Chalasani N, Younossi Z, Lavine JE, et al. The diagnosis and management of nonalcoholic fatty liver disease: practice guidance from the American Association for the Study of Liver Diseases. Hepatology. 2018;67(1):328-357. doi:10.1002/hep.29367
