Performance of Algorithms for Identifying Patients With Chronic Hepatitis B or C Infection in the French Health Insurance Claims Databases Using the ANRS CO22 HEPATHER Cohort

Laurent Lam, Hélène Fontaine, Nathanael Lapidus, Jonathan Bellet, Clovis Lusivika-nzinga, Jérôme Nicol, Céline Dorival, Carole Cagnot, Gilles Hejblum, Stanislas Pol, et al.

Lusivika Nzinga Clovis (Orcid ID: 0000-0003-4457-4228)
Lam Laurent (Orcid ID: 0000-0003-3857-9604)

Performance of Algorithms for Identifying Patients With Chronic Hepatitis B or C Infection in the French Health Insurance Claims Databases Using the ANRS CO22 HEPATHER Cohort

Short title: Performance of HBV/HCV Algorithms in the SNDS

Laurent Lam, MD; Hélène Fontaine, MD; Nathanael Lapidus, MD, PhD; Jonathan Bellet, BS; Clovis Lusivika-Nzinga, PhD; Jérôme Nicol, PhD; Céline Dorival, PhD; Carole Cagnot, PhD; Gilles Hejblum, PhD; Stanislas Pol, MD, PhD; Marc Bourlière, MD; Fabrice Carrat, MD, PhD; for the ANRS/AFEF HEPATHER study group

Affiliations: Sorbonne Université, INSERM, Institut Pierre Louis d'Épidémiologie et de Santé Publique, IPLESP, Paris, France (Lam, Lapidus, Dorival, Bellet, Lusivika Nzinga, Nicol, Hejblum, and Carrat); Sorbonne Université, Department of Public Health, Assistance Publique-Hôpitaux de Paris, Hôpital Saint-Antoine, Paris, France (Lapidus, Carrat); Department of Hepatology, Assistance Publique-Hôpitaux de Paris, Hôpital Cochin, Paris, France (Fontaine, Pol); Université de Paris Cité, Paris, France (Pol); ANRS, Emerging Infectious Diseases, Paris, France (Cagnot) ; Department of Hepatology and Gastroenterology, Hôpital Saint Joseph, Marseille, France (Bourlière); INSERM, UMR 1252 IRD SESSTIM, Aix Marseille Université, Marseille, France (Bourlière)

Corresponding Author: Fabrice Carrat, MD, PhD, Institut Pierre Louis d'Épidémiologie et de Santé Publique, 27 rue Chaligny, 75571 PARIS CEDEX 12 FRANCE (fabrice.carrat@iplesp.upmc.fr)

Tel: +33(0)144738643

Electronic word count:

Abstract: 250 words

Manuscript: 3205 words

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/jvh.13788

References: 38 references

Tables and Figures:

Number of Main Tables: 4

Number of Main Figures: 1

Acknowledgments:

We thank participants and participating clinicians at each study site. We thank Marjorie BOUSSAC and the staff of the Caisse Nationale d'Assurance Maladie for their support in providing the SNDS data.

Conflict of interest statement: FC reports grants from INSERM-ANRS, during the conduct of the study; personal fees from Sanofi, outside the submitted work. HF reports personal fees and invitations for medical meetings from Gilead, AbbVie, Bristol-Myers Squibb, MSD and Janssen, outside the submitted work. MB reports grants and personal fees from AbbVie and Gilead, outside the submitted work, and personal fees from MSD, Janssen, Boehringher Ingelheim, Intercept and Bristol-Myers Squibb, outside the submitted work. SP has received consulting and lecturing fees from Janssen, Gilead, MSD, Abbvie, Biotest, Shinogui, Viiv, LFB and grants from Gilead and Abbvie without relation to this manuscript. LL, NL, CD, JB, CL-N, JN, CC, and GH have nothing to disclose.

Funding statement: The ANRS CO22 HEPATHER cohort is sponsored and funded by INSERM-ANRS and conducted in collaboration with Association Française pour l'Étude du Foie (AFEF). The cohort received supports from the Agence Nationale de la Recherche (19-COHO-0002), DGS (Direction Générale de la Santé), MSD, Janssen, Gilead, AbbVie, BMS, and Roche. The public/private partnership is built in total transparency through a specific contract. The pharmaceutical companies are not involved in scientific decisions. The sponsor played no role in data collection, data analysis or data interpretation. The other funding sources played no role in study design, data collection, data analysis, data interpretation, or drafting the study.

Ethical statement: The HEPATHER protocol was carried out in accordance with the Declaration of Helsinki and the French law on biomedical research and was

approved by the appropriate research ethics committee (Comité de Protection des Personnes Ile de France 3), the French National Agency for Medicines and Health Products (ANSM), and the French Data Protection Authority (CNIL).

Abbreviations: ALT, Alanine Aminotransferase; AST, Aspartate Aminotransferase; BMI, body mass index; CCAM; French common classification of medical procedures; CHB, Chronic Hepatitis B; CHC, chronic hepatitis C; CI, confidence interval; CNIL, Commission Nationale de l'Informatique et des Libertés (French Data Protection Authority); DAA, direct-acting antiviral; GGT, gamma-glutamyl transferase; HBeAg, hepatitis B e antigen; HBV, hepatitis B virus; HCV, hepatitis C virus; HR, hazard ratio; IQR, interquartile range ; ICD-10, International Classification of Diseases, 10th Revision; LTD, severe and costly long-term disease; NABM, nomenclature of medical biology acts; RNA, Ribonucleic acid; SNDS, Système National des Données de Santé (French administrative health insurance database)

Author Contributions: LL and FC had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Concept and design: LL and FC. Acquisition, analysis, or interpretation of data: All authors. Drafting of the manuscript: LL. Critical revision of the manuscript for important intellectual content: All authors. Statistical analysis: LL and FC. All authors approved the final version of the submission.

Informed Consent: Informed consent from each patient before inclusion and a specific consent for the linkage with the SNDS were obtained.

Data Availability Statement: Due to the nature of this research based on the French administrative health care database, data that support the findings of this study are not available due to ethical and legal restrictions.

Abstract:

The validity of algorithms for identifying patients with chronic hepatitis B or C virus (HBV or HCV) infection in claims databases has been little explored. The performance of 15 algorithms was evaluated. Data from HBV- or HCV-infected patients enrolled between August 2012 and December 2015 in French hepatology centres (ANRS CO22 HEPATHER cohort) were individually linked to the French national health insurance system (SNDS). The SNDS covers 99% of the French population and contains health care reimbursement data. Performance metrics were calculated by comparing the viral status established by clinicians with those obtained with the algorithms identifying chronic HBV- and HCV-infected patients. A total of 14 751 patients (29% with chronic HBV and 63% with chronic HCV infection) followed-up until December 2018 were selected. Despite good specificity, the algorithms relying on ICD-10 codes performed poorly. By contrast, the multi-criteria algorithms combining ICD-10 codes, antiviral dispensing, laboratory diagnostic tests (HBV DNA or HCV RNA detection and quantification, HCV genotyping), examinations for the assessment of liver fibrosis, and long-term disease registrations were the most effective (sensitivity 0.92, 95% CI, 0.91 to 0.93 and specificity 0.96, 95% CI, 0.95 to 0.96 for identifying chronic HBV-infected patients; sensitivity 0.94, 95% CI, 0.94 to 0.94 and specificity 0.85, 95% CI, 0.84 to 0.86 for identifying chronic HCV-infected patients). In conclusion, the multi-criteria algorithms perform well in identifying patients with chronic hepatitis B or C infection and can be used to estimate the magnitude of the public health burden associated with hepatitis B and C in France.

Keywords: Sensitivity and Specificity; Routinely Collected Health Data; Validation Study; Hepatitis B; Hepatitis C.

Introduction

An increasing number of epidemiological studies are being conducted using passively collected real-world data, such as insurance and claims databases.[1] These databases are valuable tools for assessing disease prevalence, trajectory of care, health care consumption, and real-world clinical and economic outcomes.[2] However, correctly identifying patients with specific diseases is a major challenge as these administrative health databases combine clinical and administrative data and were not originally designed for clinical observational studies. Furthermore, despite the existence of empirical algorithms to identify specific diseases in claims databases, the validity of these algorithms is often unknown due to the absence of patients with a known disease status.[3]

The economic and clinical burden associated with chronic hepatitis B (CHB) or C (CHC) is significant.[4,5] The correct estimation of the burden of disease and its reproducibility between studies are highly dependent on the accurate identification of patients.[6,7] To date, several algorithms have been proposed to identify patients with chronic hepatitis B or C infection using claims databases.[8,9] Most of the studies were not conducted in European countries and were limited to simple algorithms using mainly ICD codes (International Classification of Diseases codes).[10–13] Other studies were limited to specific subgroups or small populations.[14] In addition, several studies provided few indicators to assess the performance of the algorithms or did not provide confidence intervals to determine the uncertainty of the estimates.[12,14] Therefore, the performance of HBV and HCV algorithms and their validity should be thoroughly assessed to overcome the above-mentioned limitations.

We aimed to evaluate the performance of several algorithms for identifying patients with chronic hepatitis B or C infection using the French health insurance databases

(SNDS)[15] – all compared with the clinically established viral status of HBV- and HCV-infected patients enrolled in a French prospective cohort.

Methods

Participants and data sources

Participants were recruited between 6 August 2012 and 31 December 2015 in 32 French hepatology centres. To be enrolled in the HEPATHER cohort, HBV- or HCV-infected patients should present at least one of the following inclusion criteria: chronic HBV-infected patients defined by a positive HBsAg for at least 6 months, chronic HCV-infected patients defined by the positivity for anti-HCV antibodies for at least 6 months and a positive HCV RNA, CHB with serological remission (undetectable HBsAg and HBV DNA) or cured hepatitis C (positive anti-HCV antibodies associated to an undetectable HCV RNA or negative viremia 3 months after the end of treatment), acute hepatitis B or C. Patients with ongoing hepatitis C treatment, pregnant women, HIV-infected, and minors were not included in the HEPATHER cohort. Furthermore, patients without data linkage with the SNDS, linkage errors, or uncertain viral hepatitis status were not eligible for this study. Study procedures have been previously described.[16]

The French health insurance databases (Système National des Données de Santé–SNDS) contain anonymised data on all health care reimbursements and cover approximately 99% of the French population.[15] For example, medical procedures performed, laboratory tests (without results), dispensed medications, physician visits, and ICD-10 codes recorded at discharge from hospitalisation are recorded in the SNDS. Costly or long-term diseases (LTD) registrations are also available in the SNDS.

Informed consent was obtained from each patient before the inclusion in the HEPATHER cohort and before the linkage with the SNDS. All individual data were anonymised. The HEPATHER protocol was carried out in accordance with the

Declaration of Helsinki and the French law on biomedical research and was approved by the appropriate research ethics committee (Comité de Protection des Personnes Ile de France 3), the French National Agency for Medicines and Health Products (ANSM), and the French Data Protection Authority (CNIL).

Study design

Patients from the prospective ANRS-CO22 HEPATHER cohort[16] were individually linked to the SNDS.[15] The linkage procedure used deterministic and probabilistic approaches. The deterministic linkage procedure, which is supposed to be error-free, was based on a unique patient identifier (numéro d'inscription au répertoire - NIR) and concerned 13 705 (93%) of analysed patients. The probabilistic approach identified the remaining patients in the SNDS using the following variables: date of birth, sex, place of residence, dates of medical visits, and geographic and legal identifiers of the corresponding health facilities.

Reimbursement data in the SNDS were retrospectively analysed from 2012 to the end of 2018 to identify patients with chronic hepatitis B or C infection using several predefined algorithms. The results of the algorithms were compared to the viral hepatitis status defined at baseline in the HEPATHER cohort. Given the availability of SNDS data, patients were followed-up until 31 December 2018.

Algorithms and reference viral hepatitis status

Overall, 15 algorithms for identifying patients with chronic hepatitis B or C infection were selected and applied to the SNDS data from 2012 to 2018. Algorithms are described in detail in Table 1 and Table 2. They could include ICD-10 codes in hospital claims data, laboratory diagnostic tests, medical procedures, LTD registrations, and medications indicated for HBV- and HCV-infected patients (Supplementary Table S1). The algorithms were adapted to the SNDS data (e.g.,

conversion of ICD-9 to ICD-10 codes) and originated from studies performed on claims databases, the national public health agency, or the authors themselves. The viral hepatitis status obtained with the algorithms was compared to the reference standard defined as the investigator-reported chronic viral hepatitis status at baseline in the HEPATHER cohort - this reference standard being confirmed by additional virological data (e.g., HBsAg for HBV-infected patients, HCV RNA for HCV-infected patients). Specifically, patients in the cohort diagnosed with CHC were used as a reference to assess the sensitivity of the CHC algorithms and the specificity of the algorithms for identifying chronic HBV-infected patients. Patients in the cohort diagnosed with chronic hepatitis B infection were used to estimate the sensitivity of the algorithms for identifying chronic HBV-infected patients and the specificity of the CHC algorithms. Similarly, patients with resolved or acute HBV or HCV infection were used as a reference to determine the specificity of the algorithms for detecting chronic HBV or HCV-infected patients. The performance of the algorithms developed for identifying CHC patients who received antivirals for HCV (algorithms shown in Supplementary Table S2) and patients with chronic hepatitis D among HBV-infected patients (algorithm shown in Supplementary Table S3) were also calculated. Patients in the cohort diagnosed with hepatitis B infection and HCV-infected patients not treated with antivirals for HCV were used to estimate the specificity of the algorithms identifying CHC patients who received antivirals for HCV. Similarly, patients in the cohort diagnosed with hepatitis C infection and patients with hepatitis B infection without chronic hepatitis D were used to estimate the specificity of the algorithm identifying patients with chronic hepatitis D among patients with chronic hepatitis B infection.

Statistical methods

The sample size was determined by assuming an expected sensitivity and specificity of 0.90 and 0.95, respectively, and a minimum acceptable sensitivity and specificity of 0.88 and 0.93, respectively. To achieve 90% statistical power with nominal alpha set to 0.05, the required sample size was 3163 patients with chronic viral hepatitis (B or C) and 1826 individuals without the disease.[17,18]

Patient characteristics at baseline and during follow-up were described according to the viral hepatitis status defined at enrolment. The performance of the algorithms was summarised using the following indicators: sensitivity, specificity, accuracy indicator, and Youden index derived from the sensitivity and specificity values.[19] The overall accuracy of the algorithms was determined from the sensitivity and specificity values and the prevalence of chronic HCV and HBV infection in the French general population.[20,21] For HBV-infected patients, the performance of algorithms for detecting HBeAg negative chronic infection and chronic hepatitis D patients were also reported. Confidence intervals for all indicators were reported using Clopper-Pearson exact approximation.[22] For all patients analysed, no indeterminate or missing values were present for algorithm results and viral hepatitis status.

Sensitivity analyses were performed to estimate the algorithms' accuracy when applied over a period limited to five years (from 2012 to 2016), and when analyses were limited to patients linked to the SNDS with the deterministic approach only. All statistical analyses were performed with R version 4.2.0 (Core Team, 2021).

Results

Participants

Between 6 August 2012 and 31 December 2015, 20 353 patients were recruited in the HEPATHER cohort and assessed for eligibility (Figure 1). Of these, 14 751 eligible patients had individual data linkage with the SNDS and were analysed. The median age of the patients was 54.4 years (IQR 45.9-62.8), and 8643 (59 %) were male. Overall, 4311 (29%) and 9276 (63%) were chronic HBV-infected and HCV-infected patients, respectively (Table 3). The remaining included 71 (0.5%) individuals with HBV and HCV coinfections, 1069 (7%) patients with resolved HBV or HCV infection at enrolment, and 24 (0.2%) patients with acute HBV or HCV infection. Of the 9276 CHC patients, 7950 (86%) initiated antivirals for HCV after inclusion. Among chronic HBV-infected patients, 1503 (35%) had HBeAg negative chronic infection. Additional clinical and demographic characteristics are reported in Table 3. Advanced fibrosis stage (F3-F4) was observed in 677 (16%) chronic HBV-infected patients and 4497 (48%) CHC patients. Among chronic HBV-infected individuals,162 (4%) were infected with the hepatitis D virus.

Algorithms for identifying patients with chronic hepatitis B infection

The performance of six different algorithms for identifying patients with chronic hepatitis B infection is presented in Table 4. The three algorithms based solely on the occurrence of ICD-10 codes in hospital claims data (algorithms 1 to 3) had high specificity values (between 0.98; 95% CI, 0.98 to 0.99, and 1.00; 95% CI, 0.99 to 1.00), although the sensitivity was low (between 0.15; 95% CI, 0.14 to 0.16, and 0.37; 95% CI, 0.35 to 0.38). The algorithm 4 combining reimbursements for the detection and quantification of HBV DNA and the dispensing of specific medications for HBV-infected patients had similar specificity values (0.99; 95% CI, 0.99 to 1.00)

and higher sensitivity (0.46; 95% CI, 0.45 to 0.48). According to Youden's index, the algorithm 6, which included several criteria such as ICD-10 codes recorded in hospital claims data and rehabilitation services, laboratory diagnostic tests, medical procedures performed, dispensed medications, and LTD registrations, performed best in identifying patients with chronic hepatitis B infection (sensitivity 0.92; 95% CI, 0.91 to 0.93, specificity 0.96, 95% CI 0.95 to 0.96). Notably, this algorithm accurately detected 87% (95% CI, 86% to 89%) and 95% (95% CI, 94% to 96%) of HBeAg negative chronic infection and CHB patients, respectively (Supplementary Table S4). The performance of the algorithm identifying hepatitis D patients among chronic HBV-infected patients is presented in Supplementary Table S5.

Algorithms for identifying CHC patients

The algorithms 1 to 3 (Table 4), which included only ICD-10 codes with varying numbers of occurrences in the SNDS, had moderate sensitivity (between 0.26 CI 95% 0.25 to 0.27, and 0.51 CI 95% 0.50 to 0.52) and high specificity (between 0.89, CI 95% 0.88 to 0.90, and 0.96 CI 95% 0.95 to 0.96). Compared to the first three algorithms, the more stringent algorithms (algorithms 5 and 6), requiring ICD-10 codes in combination with the reimbursement of laboratory diagnostic tests (HCV RNA detection and quantification), had even lower sensitivity (between 0.10; 95% CI, 0.10 to 0.11 and 0.15, CI 95% 0.14 to 0.16), despite a slight increase in specificity (between 0.98, 95% CI, 0.97 to 0.98, and 0.99; 95% CI, 0.98 to 0.99). Of all the algorithms identifying CHC patients, the multi-criteria algorithms (algorithms 7 and 8) performed best (sensitivity 0.94, 95% CI, 0.94 to 0.94, specificity 0.85, 95% CI, 0.84 to 0.86) and were based on a combination of ICD-10 codes in hospital claims data, dispensing of medications for HCV infection, laboratory diagnostic tests for hepatitis C (HCV RNA detection and quantification), LTD registrations, and examinations for

the assessment of liver fibrosis. In an additional analysis, the best performing algorithm for detecting CHC patients treated with antivirals for HCV (Supplementary Table S2) had a sensitivity of 0.79, 95% CI, 0.78 to 0.80 and a specificity of 0.96, IC95% 0.96 to 0.97 (Supplementary Table S6).

Sensitivity analyses

In a sensitivity analysis limiting the period over which the algorithms were applied to five years, similar performance was found for all the above-mentioned algorithms (Supplementary Table S7). Similarly, after restricting the analyses to the patients linked to the SNDS with the deterministic approach only, we found consistent results with our main findings (Supplementary Table S8).

Discussion

According to the analysis of 14 751 patients individually linked to the French health insurance databases, three main results were observed. First, the multi-criteria algorithms based on a combination of ICD-10 codes, medical procedures, laboratory tests, dispensed medications, and registrations for long-term diseases, accurately detected patients with chronic hepatitis B infection and CHC patients in the administrative claims database. Second, despite their high specificity, ICD-10 coding algorithms based on hospital discharge codes, had limited performance in identifying patients with chronic hepatitis B infection and CHC patients. Third, the algorithms were able to adequately identify specific subgroups, such as HBeAg negative chronic infection and CHC patients treated with antivirals for HCV.

Algorithms based exclusively on the occurrence of ICD-10 codes had low sensitivity, despite reasonable specificity. Of these, algorithms requiring multiple occurrences of ICD-10 codes to identify patients with chronic hepatitis B infection and CHC patients had higher specificity values. This high specificity is reassuring, as it reveals the accuracy of the diagnostic codes recorded for hospital discharge in the SNDS.

Several studies have already evaluated algorithms using ICD-9 and ICD-10 for identifying patients with chronic viral hepatitis.[10–14] In agreement with our results, a recent study using Ontario administrative data found moderate sensitivity values (12.8% and 30.8% for HBV and HCV diagnostic codes, respectively), as well as high specificity, after testing the validity of HBV and HCV diagnostic codes.[23] This study highlighted the need for algorithms that include criteria other than ICD-10 codes to better monitor the burden associated with patients with chronic hepatitis B or C infection. In another study conducted on electronic health records from four health care systems in the USA, higher sensitivity values were reported after testing

algorithms based on ICD-9 codes.[13] The intrinsic characteristics of claims databases, which are country-specific, and discrepancies between ICD-9 and ICD-10 codes for defining patients with chronic hepatitis B or C infection,[24] could explain these variations. In France, a large number of patients with chronic hepatitis B or C infection are not diagnosed during hospitalisation and are not reported in the SNDS via ICD-10 codes. In addition, because chronic hepatitis B or C infections are mostly asymptomatic, these infections may be missed by patients and clinicians during non-liver-related hospitalisations. Furthermore, ICD-10 codes associated with higher reimbursement than those specific to hepatitis B and C diagnoses may have been preferred, resulting in false-negative results. Prevalence and burden estimates based on the algorithms relying only on ICD-10 codes must therefore be corrected in the absence of comprehensive laboratory results linked to administrative healthcare data, as is the case for the British Columbia Hepatitis Testers Cohort or the Veteran Affairs Corporate Data Warehouse.

The performance of algorithms that include multiple criteria, such as laboratory diagnostic tests, medical procedures, and dispensed medications to identify patients with chronic hepatitis B or C infection, has never been validated, although they have been used to identify HBV- and HCV-infected patients from the SNDS. Several factors may explain the lack of studies validating multi-criteria algorithms in France, such as: 1) the absence of a reference standard;[3] 2) the recent access to administrative claims databases for academic researchers in France;[25,26] 3) the complexity of performing individual data linkage between external databases and claims databases;[27] 4) the technical expertise required for the analysis of administrative claims databases.[28] In a recent study describing the cascade of care

of CHC patients,[8] the authors highlighted the necessity to validate the algorithm they had developed using the SNDS, which was also included in our study.

Compared to the algorithms based solely on ICD-10 codes to identify patients with chronic hepatitis B or C infection, several multi-criteria algorithms demonstrated improved sensitivity with no decrease in specificity. Interestingly, the addition of LTD registrations to ICD-10 codes had a limited impact on the sensitivity of algorithms identifying patients with chronic hepatitis B infection. This result may be related to the under-reporting of chronically HBV-infected patients without therapeutical indication to the French health insurance agency.[29] A report from the French health insurance agency mentioned the potential beneficial impact of adding serum HBV DNA assays to the algorithms identifying HBV-infected patients.[30] Our results confirmed this hypothesis, as the inclusion of this laboratory diagnostic test in algorithms increased sensitivity by more than 30% with no reduction in specificity.

Approximately 20% of antivirals dispensed for CHC were not found in the SNDS. Several factors explain this result. Direct-acting antivirals (DAAs) were initially available only through the temporary authorisation for use (ATU) programme in France, and ATUs issued for a single named person were not systematically recorded in the SNDS.[31] In addition, coding errors precluded the record of some DAAs during the early years of DAAs availability.[32] Furthermore, some patients received DAAs before the universal access to DAAs in France or as part of clinical research protocols that could not be identified in the SNDS. As many DAAs are now also dispensed in pharmacies, they are now better recorded in the SNDS.[33] It is therefore expected that the actual performance of the algorithms for identifying CHC patients treated with antivirals will be higher.

So far, estimates of the prevalence of patients with chronic hepatitis B or C infection in France have been based mainly on surveys of at-risk populations and LTD registrations.[34–36] Consequently, the assessment of the public health burden associated with viral hepatitis based on these prevalence data might be misleading. In this study, the performance of the algorithms was assessed using a very large number of patients enrolled in several hepatology centres in France. As a result, very precise estimates and narrow confidence intervals were reported for all performance indicators. The use of validated algorithms with known performance for identifying patients with chronic hepatitis B or C infection should allow for adjustment of the burden estimates obtained from the analyses of claims databases.

Our study has several limitations. First, the reference standard groups were composed of only HBV- and HCV-infected patients and were not representative of the general population. Nevertheless, as the clinical characteristics of HBV- and HCV-infected patients are very similar, our results can be considered conservative. The performance of our algorithms, and in particular the specificity, should therefore be higher if the algorithms were applied to the general population of the SNDS. Second, HBV- and HCV-infected patients without HIV infection were recruited in tertiary care facilities. Nevertheless, a possible spectrum bias would have a limited influence on the extrinsic validity and generalisability of our results given the representativeness of HBV- and HCV-infected patients included in our study. Finally, some algorithms could include country-specific criteria, which do not exist in other claims databases. Consequently, the validity of the algorithms should be further examined in other countries.

In conclusion, for the first time, our work explored the validity and performance of algorithms for identifying patients with chronic hepatitis B or C infection using real-world data from the French administrative claims databases. The multi-criteria algorithms performed best and were based on a combination of ICD-10 codes, medical procedures, laboratory diagnostic tests, dispensed medications, and long-term illness registrations. Estimates of the clinical and economic burden of hepatitis B and C should be interpreted considering the performance of algorithms for identifying patients with chronic hepatitis B or C infection.

References

1. Pol S, Fouad F, Lemaitre M, et al. Impact of extending direct antiviral agents (DAA) availability in France: an observational cohort study (2015-2019) of data from French administrative healthcare databases (SNDS). *Lancet Reg Health - Eur*. 2022;13:100281. doi:10.1016/j.lanepe.2021.100281
2. Johnson EK, Nelson CP. Values and pitfalls of the use of administrative databases for outcomes assessment. *J Urol*. 2013;190(1):17-18. doi:10.1016/j.juro.2013.04.048
3. Koram N, Delgado M, Stark JH, Setoguchi S, Luise C. Validation studies of claims data in the Asia-Pacific region: A comprehensive review. *Pharmacoepidemiol Drug Saf*. 2019;28(2):156-170. doi:10.1002/pds.4616
4. Younossi Z, Park H, Henry L, Adeyemi A, Stepanova M. Extrahepatic Manifestations of Hepatitis C: A Meta-analysis of Prevalence, Quality of Life, and Economic Burden. *Gastroenterology*. 2016;150(7):1599-1608. doi:10.1053/j.gastro.2016.02.039
5. Lavanchy D. The global burden of hepatitis C. *Liver Int*. 2009;29:74-81. doi:10.1111/j.1478-3231.2008.01934.x
6. Alwan NA. Surveillance is underestimating the burden of the COVID-19 pandemic. *The Lancet*. 2020;396(10252):e24. doi:10.1016/S0140-6736(20)31823-7
7. King CH, Galvani AP. Underestimation of the global burden of schistosomiasis. *The Lancet*. 2018;391(10118):307-308. doi:10.1016/S0140-6736(18)30098-9
8. Brouard C, Pillonel J, Boussac M, et al. French hepatitis C care cascade: substantial impact of direct-acting antivirals, but the road to elimination is still long. *BMC Infect Dis*. 2020;20(1):759. doi:10.1186/s12879-020-05478-6
9. Isenhour C, Hariri S, Vellozzi C. Monitoring the hepatitis C care cascade using administrative claims data. *Am J Manag Care*. 2018;24(5):232-238.
10. Sheu MJ, Liang FW, Li ST, Li CY, Lu TH. Validity of ICD-10-CM Codes Used to Identify Patients with Chronic Hepatitis B and C Virus Infection in Administrative Claims Data from the Taiwan National Health Insurance Outpatient Claims Dataset. *Clin Epidemiol*. 2020;Volume 12:185-192. doi:10.2147/CLEP.S236823
11. Abara WE, Moorman AC, Zhong Y, et al. The Predictive Value of International Classification of Disease Codes for Chronic Hepatitis C Virus Infection

Surveillance: The Utility and Limitations of Electronic Health Records. *Popul Health Manag*. 2018;21(2):110-115. doi:10.1089/pop.2017.0004

12. Kramer JR, Davila JA, Miller ED, Richardson P, Giordano TP, El-Serag HB. The validity of viral hepatitis and chronic liver disease diagnoses in Veterans Affairs administrative databases: LIVER DISEASE CODES VALIDATION STUDY. *Aliment Pharmacol Ther*. 2007;27(3):274-282. doi:10.1111/j.1365-2036.2007.03572.x
13. Mahajan R, Moorman AC, Liu SJ, Rupp L, Klevens RM, Chronic Hepatitis Cohort Study (CHeCS) investigators*. Use of the International Classification of Diseases, 9th revision, coding in identifying chronic hepatitis B virus infection in health system data: implications for national surveillance. *J Am Med Inform Assoc JAMIA*. 2013;20(3):441-445. doi:10.1136/amiajnl-2012-001558
14. Niu B, Forde KA, Goldberg DS. Coding algorithms for identifying patients with cirrhosis and hepatitis B or C virus using administrative data: CODING ALGORITHMS FOR HBV AND HCV CIRRHOSIS. *Pharmacoepidemiol Drug Saf*. 2015;24(1):107-111. doi:10.1002/pds.3721
15. Tuppin P, Rudant J, Constantinou P, et al. Value of a national administrative database to guide public decisions: From the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France. *Rev DÉpidémiologie Santé Publique*. 2017;65:S149-S167. doi:10.1016/j.respe.2017.05.004
16. Carrat F, Fontaine H, Dorival C, et al. Clinical outcomes in patients with chronic hepatitis C after direct-acting antiviral treatment: a prospective cohort study. *The Lancet*. 2019;393(10179):1453-1464. doi:10.1016/S0140-6736(18)32111-1
17. Korevaar DA, Gopalakrishna G, Cohen JF, Bossuyt PM. Targeted test evaluation: a framework for designing diagnostic accuracy studies with clear study hypotheses. *Diagn Progn Res*. 2019;3(1):22. doi:10.1186/s41512-019-0069-2
18. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. 1. publ. in paperback. Oxford University Press; 2004.
19. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3(1):32-35. doi:10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3

20. Alberg AJ, Park JW, Hager BW, Brock MV, Diener-West M. The use of "overall accuracy" to evaluate the validity of screening or diagnostic tests. *J Gen Intern Med*. 2004;19(5):460-465. doi:10.1111/j.1525-1497.2004.30091.x
21. the 2016 Health Barometer Group, Brouard C, Saboni L, et al. HCV and HBV prevalence based on home blood self-sampling and screening history in the general population in 2016: contribution to the new French screening strategy. *BMC Infect Dis*. 2019;19(1):896. doi:10.1186/s12879-019-4493-2
22. Altman DG, ed. *Statistics with Confidence: Confidence Intervals and Statistical Guidelines ; [Includes Disk]*. 2. ed., [Nachdr.]. BMJ Books; 2011.
23. Yasseen AS, Kwong JC, Kustra R, et al. Validating viral hepatitis B and C diagnosis codes: a retrospective analysis using Ontario's health administrative data. *Can J Public Health*. 2021;112(3):502-512. doi:10.17269/s41997-020-00435-x
24. Topaz M, Shafran-Topaz L, Bowles KH. ICD-9 to ICD-10: evolution, revolution, and current debates in the United States. *Perspect Health Inf Manag*. 2013;10:1d.
25. Looten V, Simon M. Impact Analysis of the Policy for Access of Administrative Data in France: A Before-After Study. *Stud Health Technol Inform*. 2020;270:1133-1137. doi:10.3233/SHTI200339
26. *Décret N° 2021-848 Du 29 Juin 2021 Relatif Au Traitement de Données à Caractère Personnel Dénommé « système National Des Données de Santé »*.; 2021.
27. Harron K, Dibben C, Boyd J, et al. Challenges in administrative data linkage for research. *Big Data Soc*. 2017;4(2):205395171774567. doi:10.1177/2053951717745678
28. Moore N, Blin P, Lassalle R, Thurin N, Bosco-Levy P, Droz C. National Health Insurance Claims Database in France (SNIRAM), Système Nationale des Données de Santé (SNDS) and Health Data Hub (HDH). In: Sturkenboom M, Schink T, eds. *Databases for Pharmacoepidemiological Research*. Springer Series on Epidemiology and Public Health. Springer International Publishing; 2021:131-140. doi:10.1007/978-3-030-51455-6_10
29. Premiers états generaux de l'hépatite B. Published online 2021. https://afef.asso.fr/wp-content/uploads/2021/03/21-02-16-Premiers-Etats-Generaux-Hepatite-B-SYNTHESE-67p.pdf

30. Quantin C. Etude des algorithmes de definition de pathologies dans le Système National d'Information Interrégimes de l'Assurance Maladie (SNIIRAM). Published online 2018. https://www.ameli.fr/sites/default/files/2014_etude-algorithmes-definition-pathologies-partie-1_cartographie.pdf
31. Lelièvre N, Bruxelles J. Autorisation temporaire d'utilisation nominative d'un médicament et obligations du médecin prescripteur. *Douleurs Eval - Diagn - Trait*. 2007;8(3):182-186. doi:10.1016/S1624-5687(07)88817-2
32. Dessauce C, Rudant J, Expert A, Barthélemy P, Cadier B. Les antiviraux à action directe (AAD) dans le traitement de l'hépatite C: retour sur 18 mois de prise en charge par l'Assurance Maladie. Points de repère. 2016; 44: 1-7.
33. Pol S, Lair-Mehiri L, Vallet-Pichard A. Is elimination of HCV realistic by 2030: France. Aghemo A, ed. *Liver Int*. 2021;41(S1):45-49. doi:10.1111/liv.14862
34. Weill-Barillet L, Pillonel J, Semaille C, et al. Hepatitis C virus and HIV seroprevalences, sociodemographic characteristics, behaviors and access to syringes among drug users, a comparison of geographical areas in France, ANRS-Coquelicot 2011 survey. *Rev DÉpidémiologie Santé Publique*. 2016;64(4):301-312. doi:10.1016/j.respe.2015.10.003
35. Cadet-Taïrou A., Saïd S. and Martinez M. Profils et pratiques des usagers des CAARUD en 2012. *Tendances*. 98. https://www.ofdt.fr/publications/collections/periodiques/lettre-tendances/profils-et-pratiques-des-usagers-des-caarud-en-2012-tendances-98-janvier-2015/
36. Richard JB, Gautier A, Guignard R, Léon C, Beck F. Méthode d'enquête du Baromètre santé 2014. Saint-Denis : Institut national de prévention et d'éducation pour la santé. Published online 215AD.
37. Kim H seok, Yu X, Kramer J, et al. Comparative performance of risk prediction models for hepatitis B-related hepatocellular carcinoma in the United States. *J Hepatol*. 2022;76(2):294-301. doi:10.1016/j.jhep.2021.09.009
38. Caisse Nationale d'Assurance Maladie (CNAM), Direction de la Stratégie, des Etudes et des Statistiques - Département des Etudes sur les Pathologies et les Patients. Méthodologie médicale de la cartographie des pathologies et des dépenses, version G7 (années 2012 à 2018). Published online February 17, 2020. https://assurance-maladie.ameli.fr/sites/default/files/2020_methode-reperage-pathologies_cartographie.pdf

Table 1. Algorithms for Identifying Patients With Chronic Hepatitis B Infection.

	Algorithm description	Author (reference)†
No. 1	≥ 1 ICD-10 codes B180 or B181 in hospital claims data‡	Sheu et al.[10]
No. 2	≥ 2 ICD-10 codes B180 or B181 in hospital claims data‡	Niu et al.[14]
No. 3	≥ 2 ICD-10 codes B180 or B181 in hospital claims data‡ separated by ≥ 6 months	Mahajan et al.[13]
No. 4	Qualitative and quantitative tests for HBV DNA in serum (NABM code 4120) followed by the dispensing of entecavir or tenofovir based medications	Kim et al.[37]
No. 5	≥ 1 criteria among: - LTD with ICD-10 codes B180 or B181 and/or - ICD-10 codes B180 or B181 in hospital claims data‡ and/or - ICD-10 codes B180 or B181 in follow-up care and rehabilitation services claims data§	New algorithm
No. 6	≥ 1 criteria among: - LTD with ICD-10 codes B180 or B181 - ICD-10 codes B180 or B181 in hospital claims data‡ and/or - ICD-10 codes B180 or B181 in follow-up care and rehabilitation services claims data§ and/or - ≥ 3 claims for qualitative and quantitative tests for HBV DNA in serum (NABM code 4120) and/or - ≥ 3 dispenses at different dates of medications indicated for chronic hepatitis B¶	New algorithm

Abbreviation: ALT, Alanine Aminotransferase; AST, Aspartate Aminotransferase; DNA, deoxyribonucleic acid; GGT, gamma-glutamyl transferase; HBeAg, hepatitis B e antigen; HBV, hepatitis B virus; ICD-10, International Classification of Diseases, 10th Revision; LTD, severe and costly long-term disease; NABM: nomenclature of medical biology acts; SNDS, French administrative health insurance database.

† Reference with similar algorithm.

‡ main, related or associated diagnosis (DP, DR, DA).

§ For follow-up care and rehabilitation services (SSR), ICD-10 codes recorded as main disease manifestation [manifestation morbide principale (MMP)], etiological condition [affection étiologique (AE)] or associated diagnosis [diagnostic associé (DA)].

¶ ATC, CIP and UCD codes provided in the supplementary material

Table 2. Algorithms for Identifying Chronic Hepatitis C patients.

	Algorithm description	Author (reference)†
No. 1	≥ 1 ICD-10 codes B182 in hospital claims data‡	Sheu et al.[10]
No. 2	≥ 2 ICD-10 codes B182 in hospital claims data‡	Niu et al.[14]
No. 3	≥ 2 ICD-10 codes B182 in hospital claims data‡ separated by ≥ 6 months	Abara et al.[11]
No. 4	≥ 2 laboratory diagnostic tests for HCV RNA detection and quantification (NABM 4124) separated by ≥ 6 months	Abara et al.[11]
No. 5	Laboratory diagnostic tests for HCV RNA detection and quantification (NABM 4124) followed by ≥ 2 ICD-10 codes B182† in hospital claims data separated by ≥ 60 days	Isenhour et al.[9]
No. 6	Laboratory diagnostic tests for HCV RNA detection and quantification (NABM 4124) followed by ≥ 3 ICD-10 codes B182† in hospital claims data	Isenhour et al.[9]
No. 7	≥ 1 major criteria OR intermediate criterion AND ≥ 1 minor criteria Major criteria: - Determination of HCV genotype (NABM 4125) and/or - Dispensing of pegylated interferon and ribavirin (with same dispensed dates)§ and/or - Dispensing of direct acting antivirals for HCV§ and/or - ICD-10 codes B182¶ in hospital claims data and/or - LTD with ICD-10 code B182 and/or - ≥ 3 laboratory diagnostic tests for HCV RNA detection and quantification (NABM 4124) Intermediate criterion: - ≥ 2 laboratory diagnostic tests for HCV RNA detection and quantification (NABM 4124) Minor criteria: - Examinations for the assessment of liver fibrosis: biopsy, fibroscan, fibrotest or fibrometer (NABM 1000, 1001, 1002 or CCAM HLQM002, HLBH001, HLHH001, HLHH005, HLHJ003) and/or - LTD with ICD-10 code B182 and/or - ICD-10 codes B182 in hospital claims data as associated diagnosis	Brouard et al.[8]
No. 8	≥ 1 criteria among: - New LTD registration with ICD-10 code B182 and/or - ICD-10 code B182 in hospital claims data¶ and/or - Dispensing of pegylated interferon and ribavirin (with same dispensed dates)§ and/or - Dispensing of direct acting antivirals for HCV§ and/or - Determination of HCV genotype (NABM 4125) and/or - ≥ 3 laboratory diagnostic tests for HCV RNA detection and quantification (NABM 4124) and/or - ≥ 2 laboratory diagnostic tests for HCV RNA detection and quantification (NABM 4124) and: o ≥ 1 examinations for the assessment of liver fibrosis: biopsy, fibroscan, fibrotest or fibrometer (NABM 1000, 1001, 1002 or CCAM HLQM002, HLBH001, HLHH001, HLHH005, HLHJ003) and/or o active LTD with ICD-10 code B182 and/or o ICD-10 code B182 in hospital claims data‡	Santé Publique France[38]
No. 9	All active long-term disease (LTD) with ICD-10 B182 in addition to algorithm 8	New algorithm

Abbreviation: CCAM; French common classification of medical procedures; HCV, hepatitis C virus; ICD-10, International Classification of Diseases, 10th Revision; LTD, severe and costly long-term disease; NABM, nomenclature of medical biology acts; RNA, Ribonucleic acid; SNDS, French administrative health insurance database.

† Reference with similar algorithm.

‡ Main, related or associated diagnosis (DP, DR, DA).

§ ATC, CIP and UCD codes provided in the supplementary material.

¶ Main or related diagnosis (DP or DR).

Table 3. Characteristics of Patients Included in Analyses (n = 14 751)†.

	Chronic HBV-infected patients (n = 4311)		Chronic HCV-infected patients (n = 9276)		Chronic HBV/HCV co-infection (n = 71)		Acute or resolved HBV/HCV (n = 1093)	
		%	n	%		%		%
Age, median (IQR), years	44.2 (34.8, 56.1)		56.3 (50.4, 64.4)		56.8 (47.0, 63.1)		58.0 (50.8, 65.8)	
Sex (%)								
Female	1578	37	4100	44	30	42	400	37
Male	2733	63	5176	56	41	58	693	63
BMI (%), kg/m^2								
< 18.5	132/4269	3	303/9227	3	6/69	9	19/1083	2
≥ 18.5 to < 24.9	2066/4269	48	4696/9227	51	38/69	55	464/1083	43
≥ 25 to < 30	1449/4269	34	3015/9227	33	20/69	29	391/1083	36
≥ 30	622/4269	15	1213/9227	13	5/69	7	209/1083	19
Geographical origin (%)								
France and Europe	1485	34	7312	79	38	54	839	77
Asia	640	15	197	2	4	6	45	4
Africa	1789	42	1351	15	22	31	147	13
Other	387	9	416	5	7	10	62	6
Fibrosis scoring (%)								
F0, F1 or F2	2839/3516	81	4097/8594	48	26/61	43	397/872	46
F3	236/3516	7	1161/8594	14	11/61	18	93/872	11
F4	441/3516	13	3336/8594	39	24/61	39	382/872	44
HBeAg negative chronic infection	1503	35			29/41	41		
Hepatitis D infection	162/4255	4	-		2	3	-	
Time since HCV diagnosis (IQR), years	-		14.4 (7.1, 19.9)		12.0 (5.6, 19.9)		-	
Missing data	-		232		4		-	
Treatment for HCV during follow-up‡	-		7950	86	61	86	-	
Treatment with DAAs during follow-up‡	-		7925	85	61	86	-	

Abbreviation: BMI, body mass index; CHC, chronic hepatitis C ; DAA, direct-acting antiviral; HBeAg, Hepatitis B e antigen; HBV, Hepatitis B virus; HCV, Hepatitis C virus; IQR, interquartile range.

† Data are presented as median (25th percentile, 75th percentile [interquartile range]) and counts (%) for continuous and categorical variables, respectively. We provided denominators for categorical variables with missing data.

‡ Patient treated with DAAs for HCV or pegylated interferon and ribavirin between the inclusion date and the end of follow-up (December 2018).

Table 4. Performance of Algorithms Applied to the SNDS (n = 14 751)[†].

	Sensitivity (95% CI)[‡]	Specificity (95% CI)	Accuracy[§] (95% CI)	Youden index[¶] (95% CI)
Algorithms for identifying chronic HBV-infected patients				
No. 1	0.37 (0.35, 0.38)	0.98 (0.98, 0.98)	0.98 (0.98, 0.98)	0.35 (0.33, 0.37)
No. 2	0.18 (0.17, 0.19)	0.99 (0.99, 1.00)	0.99 (0.99, 1.00)	0.18 (0.16, 0.19)
No. 3	0.15 (0.14, 0.16)	1.00 (0.99, 1.00)	1.00 (0.99, 1.00)	0.15 (0.14, 0.16)
No. 4	0.46 (0.45, 0.48)	0.99 (0.99, 1.00)	0.99 (0.99, 1.00)	0.46 (0.44, 0.48)
No. 5	0.62 (0.60, 0.63)	0.99 (0.98, 0.99)	0.99 (0.98, 0.99)	0.59 (0.57, 0.61)
No. 6	0.92 (0.91, 0.93)	0.96 (0.95, 0.96)	0.96 (0.95, 0.96)	0.88 (0.87, 0.89)
Algorithms for identifying chronic HCV-infected patients				
No. 1	0.51 (0.50, 0.52)	0.89 (0.88, 0.90)	0.89 (0.88, 0.90)	0.40 (0.38, 0.42)
No. 2	0.31 (0.30, 0.32)	0.94 (0.94, 0.95)	0.94 (0.94, 0.95)	0.25 (0.24, 0.27)
No. 3	0.26 (0.25, 0.27)	0.96 (0.95, 0.96)	0.96 (0.95, 0.96)	0.22 (0.20, 0.23)
No. 4	0.64 (0.63, 0.65)	0.91 (0.90, 0.91)	0.91 (0.90, 0.91)	0.55 (0.53, 0.57)
No. 5	0.15 (0.14, 0.16)	0.98 (0.97, 0.98)	0.98 (0.97, 0.98)	0.13 (0.12, 0.14)
No. 6	0.10 (0.10, 0.11)	0.99 (0.98, 0.99)	0.99 (0.98, 0.99)	0.09 (0.08, 0.10)
No. 7	0.94 (0.94, 0.94)	0.85 (0.84, 0.86)	0.85 (0.84, 0.86)	0.79 (0.77, 0.80)
No. 8	0.94 (0.94, 0.94)	0.85 (0.84, 0.86)	0.85 (0.84, 0.86)	0.79 (0.77, 0.80)
No. 9	0.97 (0.96, 0.97)	0.80 (0.79, 0.81)	0.80 (0.79, 0.81)	0.77 (0.75, 0.78)

Abbreviation: CI, confidence interval; SNDS, French administrative health insurance database

[†] Algorithm applied to the SNDS data from 01/01/2012 to 12/31/2018.

[‡] All confidence intervals were computed with the Clopper-Person method.

[§] Accuracy refers to the proportion of correctly classified patients. Overall accuracy = (prevalence in the French general population)(sensitivity) + (1- prevalence in the French general population)(specificity).

[¶] The Youden Index (sensitivity + specificity - 1) summarises the performance of each algorithm. A Youden index equals to 1 indicates a perfect algorithm with no false positive or false negative patients.

Figure Legends:

Figure 1. Flow Diagram of the Patients Included in the Study.

Abbreviation: HBV, hepatitis B virus; HCV, hepatitis C virus; SNDS, French administrative health insurance database.