Published in final edited form as:

Pharmacoepidemiol Drug Saf. 2013 January ; 22(1): 103–107. doi:10.1002/pds.3367.

Validation of a coding algorithm to identify patients with hepatocellular carcinoma in an administrative database

David S. Goldberg, MD, MSCE^{1,2}, James D. Lewis, MD, MSCE^{1,2,3}, Scott D. Halpern, MD, PhD^{2,3,4}, Mark G Weiner, MD^{2,3,5}, and Vincent Lo Re III, MD, MSCE^{2,6}

¹Department of Medicine, Division of Gastroenterology

²Clinical Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania

³Leonard Davis Institute of Health Economics, University of Pennsylvania

⁴Department of Medicine, Division of Pulmonary, Allergy, and Critical Care

⁵Department of Medicine, Division of General Internal Medicine

⁶Department of Medicine, Division of Infectious Diseases

Abstract

Purpose—International Classification of Disease, Ninth Revision, Clinical Modification (ICD-9-CM) based algorithms to identify patients with hepatocellular carcinoma (HCC) have not been developed outside of the VA healthcare setting. The development and validation of such algorithms are necessary for the conduct of population-based studies evaluating the epidemiology and comparative effectiveness and safety of therapies for HCC.

Methods—We queried electronic medical records at two tertiary care hospitals to identify patients with two ICD-9-CM diagnosis codes for a chronic liver disease and/or cirrhosis plus two ICD-9-CM codes for HCC. We determined the positive predictive value (PPV) of this algorithm by comparing it to diagnoses of HCC confirmed by expert medical record review.

Results—Among 101 patients meeting the algorithm, 88 (PPV: 87.1%; 95% CI: 79.0–93.0%) had confirmed HCC. The algorithm's sensitivity was 91.7% among patients with confirmed HCC, and its specificity was 98.7% among chronic liver disease patients without HCC. Excluding patients who received systemic chemotherapy in the 12 months prior to or 6 months after the initial ICD-9-CM code in the algorithm, the PPV increased to 91.6% (87/95; 95% CI: 84.1–96.3%).

Conclusions—The presence of at least two ICD-9-CM codes for a chronic liver disease and/or cirrhosis plus two ICD-9-CM codes for HCC has a high PPV for identifying HCC cases. This simple, claims-based algorithm can be used in future epidemiologic studies to examine risk factors for HCC and evaluate outcomes and adverse events of medical therapies prescribed for HCC patients.

Keywords

Validation; hepatocellular carcinoma; chronic liver disease; cirrhosis; ICD-9-CM code

Corresponding Author: David Goldberg, Hospital of the University of Pennsylvania, 3400 Spruce Street, 9 Penn Tower, Philadelphia, PA 19104, Phone: 646-242-6349, Fax: 215-349-5915, david.goldberg@uphs.upenn.edu.

Disclosures

The authors of this manuscript have no conflicts of interest to disclose as described by Pharmacoepidemiology and Drug Safety.

Introduction

Hepatocellular carcinoma (HCC) is the sixth most common cancer, with an estimated 20,000 new cases diagnosed in the U.S. each year.(1) Population-based studies of the epidemiology and treatment outcomes for HCC have predominantly been conducted using Surveillance Epidemiology and End Results (SEER) data.(2–4) In these studies, HCC was confirmed using tissue pathology and radiology reports submitted to SEER registries.(4) However, for studies requiring data not available in SEER, the options have been limited. For example, some studies have used SEER-Medicare linkages, but this approach leads to selected samples of patients almost entirely over age 65 years and with drug data limited to beneficiaries with Medicare Part D.(4–6)

Other population-based HCC studies have been conducted among U.S. veterans using ICD-9-CM codes to identify HCC cases.(3, 7, 8) The algorithm developed and validated in the Veterans Affairs (VA) administrative database required only a single ICD-9-CM code for HCC (155.0) and had a positive predictive value (PPV) of 86%(3, 7, 8). However, because the VA relies on provider coding for outpatient diagnoses, the validity of this algorithm outside of the VA setting is unclear.

Given the importance of population-based studies evaluating the epidemiology and effectiveness of treatment for HCC among patients outside of the SEER-Medicare and VA populations, we sought to determine the ability of ICD-9-CM diagnostic codes to identify HCC patients in an administrative database.

Methods

Study design and data source

We conducted a cross-sectional study among patients cared for at two hospitals in the University of Pennsylvania Health System (UPHS).(9) We used the Pennsylvania Integrated Clinical and Administrative Research Database (PICARD), a warehouse of patient data that includes ICD-9-CM codes, laboratory test results, and ambulatory electronic health records.(9)

Study Subjects and Coding Algorithm Derivation

We queried the PICARD database to identify a random sample of 120 patients 18 years of age who had: 1) two ICD-9-CM codes (inpatient or outpatient) for HCC recorded between January 1, 1997 and December 31, 2011, and 2) two ICD-9-CM codes (inpatient or outpatient) for a chronic liver disease and/or cirrhosis (see list of codes in Table 1). Patients were required to have had an ICD-9-CM code for a chronic liver disease within 365 days of receiving a code for HCC.

We devised this algorithm based on a review of the literature and discussions with clinical experts in the field. Two prior studies evaluating the PPV of the ICD-9-CM code for HCC (155.0) had been conducted in the VA population. These studies showed that the PPV of the ICD-9-CM code for HCC was low among the general VA population (10), but high among those with chronic hepatitis C virus infection.(3) Given these findings, and since HCC occurs almost exclusively in the setting of chronic liver disease, we sought to develop a coding algorithm that required both an ICD-9-CM diagnostic code for HCC and a chronic liver disease (including cirrhosis).(11, 12) Second, given the complexity of the HCC and chronic liver disease diagnoses clinically, and prior work demonstrating the increased PPV with the use of two ICD-9-CM codes,(10) we required that two ICD-9 codes be recorded for each condition. This served to increase the PPV, and maximize the specificity, of our algorithm. Finally, during medical record review of the identified cases, we noted that those

who received systemic chemotherapy in the 6 months after receiving an ICD-9-CM code for HCC often did not have HCC after chart adjudication. As a result, we added the final provision for systemic chemotherapy as a secondary analysis (see Data Analysis section). We estimated that a sample of 100 patients would allow determination of the PPV within a confidence interval width of $\pm 8\%$, assuming a PPV of 80%. We oversampled and requested 120 records, assuming 15% of the charts would not have sufficient information available for review.

Outcomes

The primary outcome was clinically confirmed HCC, defined by radiographic and/or histologic criteria.(1, 11, 13, 14) The radiographic diagnosis was confirmed by reviewing all reports of contrast-enhanced CT and/or MRI scans. HCC was defined by: 1) a contrast-enhanced CT or MRI demonstrating a liver mass ≥ 1 cm with arterial hypervascularity and venous or delayed phase washout in the setting of a chronic liver disease, consistent with published guidelines(13); or 2) a biopsy of a liver mass confirming HCC. The presence of a chronic liver disease and/or cirrhosis was confirmed by serologic confirmation and/or documentation in the medical record.

As secondary outcomes, we evaluated whether patients with confirmed HCC met Milan or University of California-San Francisco (UCSF) criteria, used to HCC amenable to cure with liver transplantation, with minimal risk of post-transplant recurrence.(1) Milan criteria is defined as one tumor ≤ 5.0 cm or up to 3 tumors, with none larger than 3.0 cm, and no macrovascular (i.e. portal vein) invasion. UCSF criteria is defined as one tumor ≤ 6.5 cm or up to 3 tumors, with none larger than 4.5 cm, with the total tumor diameter less than 8 cm, and no macrovascular invasion.(1) The determination of whether a patient was within transplant criteria was determined on the first date the patient received an ICD-9-CM code for HCC.

Medical records were reviewed by a single hepatologist (D.G.) to confirm outcomes.

Data Analysis

We determined the PPVs of our algorithm to identify: a) HCC; b) HCC meeting Milan criteria; and c) HCC meeting UCSF criteria (Table 1). We also evaluated the PPV of the HCC algorithm stratified by sex and age group (>60 years versus ≤ 60 years, as 60 was the median age patients received their first HCC code). Although we did not use pharmacy data to identify our cohort, we predicted that PPV might be increased in future studies using pharmacy data to exclude patients receiving chemotherapy (other than sorafenib, used in HCC(13)). Our reasoning was that patients with metastases to the liver or with pancreatic or biliary malignancies may be miscoded as HCC, and these patients could be identified and appropriately excluded by determining utilization of chemotherapy. Thus, we performed a secondary analysis in which we determined the PPV of our algorithm after excluding patients who on medical record review received any systemic chemotherapy (excluding sorafenib) in the 12 months prior to or 6 months after their initial HCC ICD-9-CM code.

We focused on PPV because if this parameter is sufficiently high, researchers can have confidence that the algorithm will identify future patient samples with high probabilities of having true HCC. We also estimated the algorithm's sensitivity by randomly selecting 60 patients with confirmed liver disease and HCC ("true positives") among 6,000 patients evaluated for liver transplantation at UPHS, and determined the proportion that met the coding algorithm. We estimated specificity by randomly selecting 60 patients with confirmed liver disease without HCC ("true negatives") from the same sample of 6,000 patients, and determined the proportion of non-HCC cases that did not meet the algorithm.

All data were analyzed using Stata 12.0 (Stata Corp, College Station, TX, USA). The study was approved by the Institutional Review Board of the University of Pennsylvania.

Results

A total of 101 patients meeting the algorithm had available records for review. Eighty-eight (87.1%; 95% CI: 79.0–93.0%) patients had confirmed HCC (Table 1), and of these, 56 (63.6%) had HCC meeting Milan criteria, and 59 (67.1%) met UCSF criteria (Table 1).

Of the 88 HCC patients, the most common etiologies of chronic liver disease were: hepatitis C (48; 54.6%), hepatitis B (10; 11.4%), non-alcoholic steatohepatitis 9; 10.2%), and cryptogenic cirrhosis (8; 9.1%). Eighty (90.9%) HCC patients had radiographic or histological evidence of cirrhosis.

The sensitivity of the primary algorithm was 91.7% (55/60; 95% CI: 81.6%–97.2%) among patients with confirmed HCC, and the specificity was 98.3% (59/60; 95% CI: 95.0%–100.0%) among those without HCC.

After excluding patients who received systemic chemotherapy, the algorithm's PPV increased to 91.6% (87/95; 95% CI: 84.1–96.3%). Of the six excluded patients, one had HCC but was receiving chemotherapy for lymphoma, while five (4 with cholangiocarcinoma, 1 with pancreatic adenocarcinoma) did not have HCC.

Although the sensitivity of the algorithm increased with increasing number of claims for HCC (median=10), the PPV remained above 80% for patients with fewer than 10 claims (39/48, 81.3%), three of whom would have been excluded with an algorithm accounting for chemotherapy (yielding PPV of 39/45, 86.7%). No statistically significant differences were observed when results were either stratified by sex (63/69 [91.3%] for males vs. 25/32 [78.1%] for females; $p=0.07$) or by age group (48/53 [90.6%] for patients > 60 years vs. 40/48 [83.3%] for those ≤ 60 years; $P=0.28$).

Discussion

This study examined the ability of ICD-9-CM codes to identify HCC patients in an administrative database. An algorithm requiring two ICD-9-CM codes for a chronic liver disease and/or cirrhosis, plus two ICD-9-CM codes for HCC, yielded a PPV of 87% for true HCC. The PPV increased to greater than 91% when we excluded patients who received systemic chemotherapy, suggesting this algorithm may be particularly useful when employed in settings in which pharmacy data are available. Since PICARD contains ICD-9-CM codes that are used for billing purposes to insurance companies, it is likely that this algorithm would perform similarly in insurance claims-based databases.

This work is important for future epidemiologic research on HCC, the incidence and prevalence of which is projected to continue to increase considerably over the next 10 years. (15, 16) While there are an increasing number of medical and interventional radiological therapies that have demonstrated a survival benefit in single- or multi-center studies, their comparative effectiveness at the population level has not been fully evaluated. Large samples of patients with valid HCC diagnoses are needed for this purpose. Additionally, such samples are needed for pharmacoepidemiologic work aimed at identifying toxicities of such therapies in patients of all age groups. Our algorithm makes such studies possible. However, while this algorithm can identify patients with HCC with a high PPV, its performance to specifically identify patients with HCC within transplant criteria is lower, and thus requires additional considerations if the goal is to identify only those patients who are potential transplant candidates.

Nonetheless, this study has limitations. First, the UPHS includes a tertiary care center offering liver transplantation and specialized oncologic care, with a referral base across Pennsylvania, Delaware, and New Jersey. It is expected there would be a large cohort of HCC patients cared for at this site, which may increase the algorithm's PPV. However, the PPV we describe is similar to that in the VA population (8). Additionally, for the same reason that we might expect a higher prevalence of HCC, we would also expect a higher prevalence of other GI malignancies that could be miscoded as HCC (i.e. cholangiocarcinoma), which would likely counterbalance the increased number of HCC cases. While the PPV for HCC meeting Milan or UCSF criteria may have been inflated due to referrals for transplant evaluation, it is also possible that there were an increased number of HCC cases outside of transplant criteria referred for clinical trials or other therapies. Second, our coding algorithm might miss HCC cases, however this is unlikely because: 1) the algorithm's negative predictive value is expected to be high given the rarity of HCC in the population; and 2) the algorithm's sensitivity exceeded 90%. Finally, the sensitivity and specificity may be limited as we defined "true positives" and "true negatives" as patients evaluated for transplantation. While this may lead to spectrum bias, the extremely high sensitivity and specificity suggest the algorithm would likely still perform well in other databases.

In conclusion, a coding algorithm that includes at least two ICD-9-CM codes for a chronic liver disease and/or cirrhosis plus two ICD-9-CM codes for HCC had a high PPV for identifying HCC patients. This algorithm can be used in future epidemiologic studies to examine the effectiveness and potential adverse events of treatments for HCC patients.

Acknowledgments

We thank Dr. Daniel Mines for his help in designing the coding algorithms to be used for this manuscript.

Financial Support

1. NIH/NIDDK F32 1-F32-DK-089694-01 Grant (DG)
2. NIH/NIAID K01 AI-070001 (VLR)
3. NIH K24-DK078228 (JL)

References

1. El-Serag HB. Hepatocellular carcinoma. *N Engl J Med*. 2011; 365(12):1118–1127. Epub 2011/10/14. [PubMed: 21992124]
2. Davila JA, Duan ZG, McGlynn KA, El-Serag HB. Utilization and Outcomes of Palliative Therapy for Hepatocellular Carcinoma A Population-based Study in the United States. *Journal of Clinical Gastroenterology*. 2012; 46(1):71–77. [PubMed: 22157221]
3. Davila JA, Henderson L, Kramer JR, Kanwal F, Richardson PA, Duan ZG, et al. Utilization of Surveillance for Hepatocellular Carcinoma Among Hepatitis C Virus-Infected Veterans in the United States. *Annals of Internal Medicine*. 2011; 154(2):85. [PubMed: 21242365]
4. Davila JA, Morgan RO, Richardson PA, Du XL, McGlynn KA, El-Serag HB. Use of Surveillance for Hepatocellular Carcinoma Among Patients With Cirrhosis in the United States. *Hepatology*. 2010; 52(1):132–141. [PubMed: 20578139]
5. Davila JA, Morgan RO, Shaib Y, McGlynn KA, El-Serag HB. Hepatitis C infection and the increasing incidence of hepatocellular carcinoma: a population-based study. *Gastroenterology*. 2004; 127(5):1372–1380. Epub 2004/11/03. [PubMed: 15521006]
6. Welzel TM, Graubard BI, Zeuzem S, El-Serag HB, Davila JA, McGlynn KA. Metabolic syndrome increases the risk of primary liver cancer in the United States: a study in the SEER-Medicare database. *Hepatology*. 2011; 54(2):463–471. Epub 2011/05/04. [PubMed: 21538440]

7. El-Serag HB, Kramer JR, Chen GJ, Duan Z, Richardson PA, Davila JA. Effectiveness of AFP and ultrasound tests on hepatocellular carcinoma mortality in HCV-infected patients in the USA. *Gut*. 2011; 60(7):992–997. Epub 2011/01/25. [PubMed: 21257990]
8. Davila JA, Weston A, Smalley W, El-Serag HB. Utilization of screening for hepatocellular carcinoma in the United States. *J Clin Gastroenterol*. 2007; 41(8):777–782. Epub 2007/08/19. [PubMed: 17700427]
9. Goldberg DLJ, Halpern SD, Weiner M, Lo Re V III. Validation of three coding algorithms to identify patients with end-stage liver disease in an administrative database *Pharmacoepidemiology and Drug Safety*. 2012 In Press.
10. Lo Re V 3rd, Lim JK, Goetz MB, Tate J, Bathulapalli H, Klein MB, et al. Validity of diagnostic codes and liver-related laboratory abnormalities to identify hepatic decompensation events in the Veterans Aging Cohort Study. *Pharmacoepidemiol Drug Saf*. 2011; 20(7):689–699. Epub 2011/06/01. [PubMed: 21626605]
11. Forner A, Llovet JM, Bruix J. Hepatocellular carcinoma. *Lancet*. 2012; 379(9822):1245–1255. Epub 2012/02/23. [PubMed: 22353262]
12. Sherman M. Epidemiology of hepatocellular carcinoma. *Oncology*. 2010; 78(Suppl 1):7–10. Epub 2010/07/17. [PubMed: 20616577]
13. Bruix J, Sherman M. Management of hepatocellular carcinoma: an update. *Hepatology*. 2011; 53(3):1020–1022. Epub 2011/03/05. [PubMed: 21374666]
14. Sherman M. The radiological diagnosis of hepatocellular carcinoma. *Am J Gastroenterol*. 2010; 105(3):610–612. Epub 2010/03/06. [PubMed: 20203642]
15. Kanwal F, Hoang T, Kramer JR, Asch SM, Goetz MB, Zeringue A, et al. Increasing prevalence of HCC and cirrhosis in patients with chronic hepatitis C virus infection. *Gastroenterology*. 2011; 140(4):1182–1188. e1. Epub 2010/12/28. [PubMed: 21184757]
16. Simard EP, Ward EM, Siegel R, Jemal A. Cancers with increasing incidence trends in the United States; 1999 through 2008. *CA Cancer J Clin*. 2012 Epub 2012/01/28.

Take-home messages

1. Population-based epidemiology studies of hepatocellular carcinoma have been limited to the SEER, SEER-Medicare, and Veterans Affairs populations.
2. ICD-9-CM-based algorithms to identify patients with HCC have not been developed outside of the VA healthcare setting.
3. Use of SEER-Medicare linkages leads to selected samples of patients almost entirely over age 65 years and with drug data limited to beneficiaries with Medicare Part D.
4. The presence of at least two ICD-9-CM codes for a chronic liver disease and/or cirrhosis plus two ICD-9-CM codes for HCC has a high PPV for identifying HCC patients.
5. This simple, claims-based algorithm therefore can be used in future epidemiologic studies to examine risk factors for HCC and evaluate outcomes and adverse events of medical therapies prescribed for HCC patients.

Table 1

Positive predictive values of International Classification of Disease, Ninth Revision, Clinical Modification (ICD-9-CM) diagnostic coding algorithms to identify patients with hepatocellular carcinoma among those with a chronic liver disease or cirrhosis* (n=101). All patients had two ICD-9-CM codes for hepatocellular carcinoma plus two ICD-9-CM codes for a chronic liver disease or cirrhosis.

Diagnosis	Number with diagnosis	PPV	95% CI of PPV
Hepatocellular carcinoma	88	87.1%	79.0–93.0%
Hepatocellular carcinoma that met Milan criteria	56	55.5%	45.2–65.3%
Hepatocellular carcinoma that met University of California-San Francisco criteria	59	58.4%	48.2–68.1%
Hepatocellular carcinoma excluding patients receiving systemic chemotherapy [†]	87	91.6%	84.1–96.3%

* ICD-9-CM codes used to identify chronic liver diseases or cirrhosis included: 070.20, 070.21, 020.22, 070.23, 070.3, 070.31, 070.32, 070.33 (hepatitis B); 070.40, 070.49, 070.59, 070.60, 070.70, 070.71, 070.90, 573.1 (viral hepatitis NOS); 070.41, 070.44, 070.51, 070.54, 070.70, 070.71 (hepatitis C); 070.42, 070.52 (hepatitis D with hepatitis B); 273.4 (alpha-1-antitrypsin deficiency); 275.0 (disorders iron metabolism); 275.1 (disorders copper metabolism); 453.0 (Budd-Chiari syndrome); 571.0, 571.1 (alcoholic liver disease); 571.40, 571.41, 571.8, 571.9, 573.0, 573.3, 573.8, 573.9 (hepatitis, liver disorders NOS); 517.42 (autoimmune hepatitis); 576.1 (cholangitis); 571.2 (alcoholic cirrhosis), 571.5 (cirrhosis without mention of alcohol, 571.6 (biliary cirrhosis). + = ICD-9-CM code present; – = ICD-9-CM code absent

[†] Total N=95, excluding patients receiving chemotherapy for cholangiocarcinoma (4) or pancreatic adenocarcinoma (1) in the absence of HCC, or lymphoma in the presence of HCC (1)