HHS Public Access

Author manuscript

Dig Dis Sci. Author manuscript; available in PMC 2017 December 07.

Published in final edited form as:

Dig Dis Sci. 2014 December ; 59(12): 3053–3061. doi:10.1007/s10620-014-3272-6.

Comparative Effectiveness Research of Chronic Hepatitis B and C Cohort Study (CHeCS): Improving Data Collection and Cohort Identification

Mei Lu,
Departments of Public Health Sciences, Center for Health Services Research, and Gastroenterology, Henry Ford Health System, One Ford Place, 3E, Detroit, MI 48202, USA

Loralee B. Rupp,
Departments of Public Health Sciences, Center for Health Services Research, and Gastroenterology, Henry Ford Health System, One Ford Place, 3E, Detroit, MI 48202, USA

Anne C. Moorman,
Division of Viral Hepatitis National Center for HIV, Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention, Atlanta, GA, USA

Jia Li,
Departments of Public Health Sciences, Center for Health Services Research, and Gastroenterology, Henry Ford Health System, One Ford Place, 3E, Detroit, MI 48202, USA

Talan Zhang,
Departments of Public Health Sciences, Center for Health Services Research, and Gastroenterology, Henry Ford Health System, One Ford Place, 3E, Detroit, MI 48202, USA

Lois E. Lamerato,
Departments of Public Health Sciences, Center for Health Services Research, and Gastroenterology, Henry Ford Health System, One Ford Place, 3E, Detroit, MI 48202, USA

Scott D. Holmberg,
Division of Viral Hepatitis National Center for HIV, Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention, Atlanta, GA, USA

Philip R. Spradling,
Division of Viral Hepatitis National Center for HIV, Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention, Atlanta, GA, USA

Eyasu H. Teshale,
Division of Viral Hepatitis National Center for HIV, Hepatitis, STD, and TB Prevention, Centers for Disease Control and Prevention, Atlanta, GA, USA

Correspondence to: Mei Lu.

A list of the CHeCS Investigators appears in Appendix 1.

Conflict of interest Stuart C. Gordon receives grant/research support from AbbVie Pharmaceuticals, Bristol-Myers Squibb, Gilead Pharmaceuticals, GlaxoSmithKline, Intercept Pharmaceuticals, Merck, and Vertex Pharmaceuticals. He is also a consultant/advisor for Amgen, Bristol-Myers Squibb, CVS Caremark, Gilead Pharmaceuticals, Merck, Novartis, and Vertex Pharmaceuticals and is on the Data Monitoring Board for Tibotec/Janssen Pharmaceuticals. The other authors have no potential conflicts of interest.

Vinutha Vijayadeva,
The Center for Health Research, Kaiser Permanente Hawaii, Honolulu, HI, USA

Joseph A. Boscarino,
Center for Health Research, Geisinger Clinic, Danville, PA, USA

Mark A. Schmidt,
The Center for Health Research, Kaiser Permanente Northwest, Portland, OR, USA

David R. Nerenz, and
Departments of Public Health Sciences, Center for Health Services Research, and Gastroenterology, Henry Ford Health System, One Ford Place, 3E, Detroit, MI 48202, USA

Stuart C. Gordon
Departments of Public Health Sciences, Center for Health Services Research, and Gastroenterology, Henry Ford Health System, One Ford Place, 3E, Detroit, MI 48202, USA

Abstract

Background and Aims—The Chronic Hepatitis Cohort Study (CHeCS) is a longitudinal observational study of risks and benefits of treatments and care in patients with chronic hepatitis B (HBV) and C (HCV) infection from four US health systems. We hypothesized that comparative effectiveness methods—including a centralized data management system and an adaptive approach for cohort selection—would improve cohort selection while controlling data quality and reducing the cost.

Methods—Cohort selection and data collection were performed primarily via the electronic health record (EHR); cases were confirmed via chart abstraction. Two parallel sources fed data to a centralized data management system: direct EHR data collection with common data elements, and chart abstraction via electronic data capture. An adaptive Classification and Regression Tree (CART) identified a set of electronic variables to improve case ascertainment accuracy.

Results—Over 16 million patient records were collected on 23 case report forms in 2006–2008. The vast majority of data (99.2 %) were collected electronically from EHR; only 0.8 % was collected via chart abstraction. Initial electronic criteria identified 12,144 chronic hepatitis patients; 10,098 were confirmed via chart abstraction with positive predictive values (PPV) 79 and 83 % for HBV and HCV, respectively. CART-optimized models significantly increased PPV to 88 for HBV and 95 % for HCV.

Conclusions—CHeCS is a comparative effectiveness research project that leverages electronic centralized data collection and adaptive cohort identification approaches to enhance study efficiency. The adaptive CART model significantly improved the positive predictive value of cohort identification methods.

Keywords

Chronic hepatitis B; Chronic hepatitis C; Comparative effectiveness research; Classification and Regression Trees; Cohort identification

Introduction

Comparative effectiveness research (CER) emphasizes the importance of conducting research in “real-world” settings as well as comparing treatment and research methods to determine “best practices” [1]. CER is used to optimize pragmatic randomized controlled trials within healthcare environments [2] while improving the rigor of observational studies to approximate the data quality of randomized clinical trials. CER methods include reduction in study costs through use of automated EHR data collection, ensuring data quality through centralized data management systems, as well as the application of “adaptive approaches” to improve study efficiency [3–5]. Both the Institute of Medicine and the American Association for the Study of Liver Disease (AASLD) support the application of CER methods to research on hepatobiliary disease [6–8]. Given the documented disparities in healthcare access and coverage among people with chronic liver disease, interventions designed to improve the quality of care will impact a significant proportion of this population [8].

The longitudinal Chronic Hepatitis B and C Cohort Study (CHeCS) was initiated in 2006 to assess the risks and benefits of hepatitis treatments and care—a CER approach. Participants were enrolled into the study from four large US health systems [9] that are members of the HMO Research Network (HMORN) [10]; data were collected primarily from the electronic health records (EHR) using a Virtual Data Warehouse (VDW). Because the existing HMORN VDW data structure uses a decentralized rather than a centralized data model—which may reduce data quality and collection efficiency [11]—the CHeCS data coordinating center (DCC) implemented centralized data collection, including additional data collection via chart abstraction. A predefined set of automated EHR-based ICD-9 codes and laboratory inclusion criteria were used to identify patients with chronic HBV and HCV at each site [9]. The chronic HBV and HCV cases were later confirmed via chart review, which was a labor intensive and costly endeavor. We hypothesized that a CER electronic initiative [1] for data collection and an adaptive approach for cohort selection would enhance study accuracy and efficiency.

Methods

Cohort Selection

The CHeCS investigation follows the guidelines of the US Department of Health and Human Services regarding the protection of human subjects. The study protocol was approved and is renewed annually by the institutional review board at each participating site. Patients were enrolled from four HMORN health systems—Henry Ford Health System (HFHS), Detroit, MI (leading clinical site and Data Coordinating Center); Geisinger Health System, Danville, PA; Kaiser Permanente–Northwest, Portland, OR; and Kaiser Permanente–Hawaii, Honolulu. For the initial enrollment cycle, the total source population included 1,248,558 adult patients with at least one health system encounter during 2006–2008 [12]. Automated EHR-based data-pulling algorithms, based on a common set of ICD-9 codes and laboratory-based EHR inclusion criteria, were used to identify patients with chronic HBV and HCV at each site. Detailed inclusion criteria have been published previously [9] and are summarized in “Appendix 2.”

CHeCS was designed to enroll new cohorts in subsequent cycles while updating existing patient data annually. However, funding was insufficient to support chart reviews for all cases identified during the second (2008–2010) and third (2011) cycles of cohort identification. Based on what we had learned from the initial cohort selection and validation, an adaptive approach was considered for future cohort identification.

Data Collection and Management

A large proportion of CHeCS data were collected specifically to assess the risks and benefits of hepatitis treatments and care in a "real-world setting." CHeCS extended and modified the existing HMORN EHR-based VDW structure [11]. The VDW structure includes a standardized database with common data elements and records at the patient level (Fig. 1) and contains automated EHR data (demographics, encounters, laboratory results), health plan data (plan enrollment, pharmacy data), and census data (e.g., estimated household income based on block group or zip code) [11]. Unlike the HMORN VDW structure, which stores data behind separate security firewalls at each health system, CHeCS data are stored in a secure, centralized data management system at HFHS with an electronic data capture (EDC) feature for direct data collection at the site, designed by the Data Coordinating Center (DCC) at HFHS.

Eleven VDW-based case report forms (CRFs) were adopted, with three additional CHeCS-specific CRFs designed in VDW-like formats (Table 1). To harmonize with the EHR data, an additional eleven CRFs were created for medical chart abstraction, including liver biopsy results, external laboratory results, and detailed antiviral therapy data from 2006 to 2010, as well as a summary history of antiviral therapy prior to 2006. These data were entered into the central HFHS database from each site via EDC. Additionally, a one-time survey of hepatitis exposure and patient behavior was mailed to participants; survey data were stripped of identifying information before being entered into the central HFHS database.

Experienced medical abstractors reviewed patients' EHR charts at each site (from first system encounter forward) for case confirmation and chart abstraction data collection. Charts were flagged if they lacked documentation that a liver specialist had diagnosed the patient with chronic hepatitis B or C, or if there was an indication of acute hepatitis or that chronic hepatitis had been ruled out. Flagged charts subsequently underwent formal case review under the supervision of a hepatitis clinician using a standardized set of hepatologist-developed criteria. Patients whose chronic hepatitis infection status could not be con-firmed were excluded [9].

Endpoints and Covariates of Interest

For this analysis, the endpoints were confirmed chronic HBV and HCV cases, respectively. The study goal was to improve cohort identification based on electronic data. Potential predictor variables (covariates), described below, included the original e-based cohort inclusion criteria (see Appendix 2), selected laboratory test results, HIV status, and 41 additional liver disease-related procedures/diagnoses:

1. The original e-based cohort inclusion criteria:

The original six HBV cohort inclusion criteria and three HCV cohort inclusion criteria detailed previously [9] and summarized in “Appendix 2.”

2. Electronic laboratory test results not included above:

   Four additional HCV variables were based on electronic laboratory results: (a) At least two positive HCV antibody test results (anti-HCV, IgG anti-HCV, or HCV RIBA); (b) at least two positive HCV antibody test results occurring at least 6 months apart; (c) at least two positive virology test results for HCV (HCV RNA qualitative, HCV RNA quantitative, or HCV RNA genotype); and (d) at least two positive HCV virology results occurring at least 6 months apart. Original HBV inclusion criteria already encompassed all laboratory tests of interest.

3. Liver-related diagnosis and procedure codes using VDW encounter codes:

   Any of the following in EHR data: (a) A biopsy procedure performed (Current Procedural Terminology [CPT] codes 47000, 47100, 47001; ICD-9 procedure codes 50.11, 50.12, 50.13, 50.14); (b) acute HBV (ICD-9 diagnosis codes 70.20, 70.21, 70.30, 70.31); (c) acute HCV (ICD-9 diagnosis codes 70.41, 70.51, 70.70, 70.71); and (d) HIV (ICD-9 diagnosis codes 42.xx, 79.53, 795.71, V08). In addition, the procedure-based and ICD-9 diagnosis-based codes for liver transplant; hepatocellular carcinoma; liver failure including hepatorenal syndrome; hepatic encephalopathy; portal hypertension (and portal decompression procedures); esophageal varices; other gastrointestinal hemorrhage (selected); ascites and paracentesis procedures; other sequelae; and indication of cirrhosis (ICD-9 diagnosis codes 571.2 and 571.5) were extracted based on electronic VDW encounter data.

HBV and HCV treatment-related variables (EHR pharmacy data) were not included as potential predictor variables in order to avoid cohort selection bias related to treatment.

Statistical Analysis

The analysis was performed based on the initial cohort data and patients (n = 12,144). All patients had undergone chart review and detailed case reviews as needed for case confirmation. Prior to the analysis, the initial cohort data were randomly divided into two datasets using SAS 9.3 [13]. One set of data—learning data—was used to build a model; the other set—testing data—was used to validate the model. Classification and Regression Trees (CART) analysis was performed using CART® 6.0 software (Sal-ford Systems, San Diego, CA) [14] to identify a set of variables to improve accuracy of cohort selection methods. CART generates a binary recursive tree partitioning, using a nonparametric approach, to identify the variables most predictive of the outcome of interest and subsequently develop a predictive model for classification of future subjects [14]. Unlike multivariable logistic regression, it is ideally suited for a clinical decision-making model because it can reveal important relationships between variables that can remain hidden when using logistic regression [15, 16].

CART begins with the root node (all subjects) and then determines which variable has the highest predictive ability to assign subjects to groups (i.e., case and control). Subsequently, it

determines the optimal split (cutoff point) for that predictor variable, which will divide the population into two child nodes with a determination of group classification (case and control) at each node, based on its prevalence. The process continues classifying the subjects until no further predictor variables are identified [14, 16].

Sensitivity, specificity, and positive predictive value (PPV), as well as model accuracy, measured by area under the receiver operating characteristic curve (AUROC), were calculated to assess the model predictive accuracy. The AUROC range is 0–1; values of 0.7–8 and 0.9 are considered “good” and “excellent” prediction, respectively. Generalized estimating equations (GEE) [17, 18] were used to compare the PPV difference between the modified inclusion criteria (CART model) and original inclusion criteria alone.

Results

Initial Cohort Selection and Confirmation

For the initial cohort, 12,144 patients were identified: 2,538 (20.9 %) met the original HBV inclusion criteria, and 9,851 (81.1 %) met the original HCV criteria, including 245 (2.0 %) who met inclusion criteria for both HBV and HCV. After chart abstraction, 10,098 patients were confirmed (1,992 HBV and 8,171 HCV, including 65 HBV/HCV co-infected). PPV rates of the original inclusion criteria were 78.5 % for HBV, 82.9 % for HCV, and 26.5 % for co-infection.

Cohort Data Collection and Management

Data were collected on a total of 26 CRFs (11 VDW-based EHR-CRFs, 3 additional EHR-CRFs, 11 abstraction CRFs, and a single CRF for survey data). During the initial data-collection cycle, 16,894,798 records were collected for the initial cohort of 12,144 HBV and HCV patients. Of the nonsurvey data records, 99.2 % were collected by automated data abstraction and 0.8 % by chart abstraction (Table 1).

Data quality (missing, inconsistent, or invalid records/data fields) was checked daily by the data coordinating center (DCC). Queries were sent back to each site monthly for resolution; after resolution, sites re-submitted the data. The DCC also harmonized data to increase efficiency compared with data processed at the site level.

Adaptive CART Models for Cohort Identification

The initial cohort (n = 12,144) was randomly divided in half, with 6,122 patients in the learning dataset, and 6,022 patients in the testing dataset. The two datasets had equal proportions of HBV and HCV, and similar patient characteristics.

Chronic Hepatitis B (HBV)

A total of 65 variables (or variable combinations) were included in the initial CART model. The final decision tree for HBV consisted of four terminal nodes (TNs): three TNs classified as HBV (TNs 2–5, Fig. 2) and one classified as non-HBV (TN1, Fig. 2). Three out of the six original inclusion criteria (see “Appendix 2”) remained in the final model. Based on the learning data, estimated PPV was 86.4 %, and overall predictive ability (AUROC) was 0.96;

sensitivity was 95.0 %, and specificity was 97.1 %. Validation results (based on the testing data) yielded a PPV of 88.1 %, AUROC of 0.97, and sensitivity and specificity of 96.0 and 97.4 %, respectively (Table 2).

The original cohort e-inclusion criteria 2, 5, and 3 (see “Appendix 2”), in hierarchical order, predicted HBV cases (Fig. 2). Criterion 2 identified 923 possible HBV cases (with 108 misclassified, TN4) as the first variable split in the model. In cases without criterion 2, a finding of two or more positive laboratory results for hepatitis B surface antigen (HBsAg), hepatitis e-antigen, or HBV DNA, in any combination occurring at least 6 months apart (criterion 3), identified an additional 95 cases (with 9 misclassified, TN3). Finally, patients without inclusion criteria 2 or 3, but with a positive laboratory result for both hepatitis B core antibody and surface antigen (criterion 5), added an additional 82 cases (33 misclassified, TN2).

Chronic Hepatitis C (HCV)

Four variables remained in the final model (Fig. 3, in hierarchical order) with five terminal nodes (TN): (1) two positive RNA test results; (2) original cohort inclusion criteria 1 (two ICD-9 codes for hepatitis C, at least 6 months apart); (3) a biopsy encounter; and 4) a cirrhosis encounter. The model had an estimated PPV of 95.6 % and AUROC of 0.93, based on the learning data. Validation results (based on the testing data) showed PPV of 94.9 % and AUROC 0.93 (Table 2). The sensitivity and specificity were 82.7 and 92.2 % based on the learning data, and 84.5 and 90.6 % based on testing data.

From the revised tree structure (Fig. 3), the two positive RNA tests identified 3406 possible cases (with 120 mis-classified, TN5) as the first variable split in the model. In patients without two positive RNA tests (TN1) or two viral hepatitis C ICD-9 codes at least 6 months apart (inclusion criterion 1), there were 1,472 controls identified (with 116 misclassified, TN1). In the absence of two positive RNA tests, the presence of the previous inclusion criterion 1 (two ICD-9 codes for hepatitis C at least 6 months apart) as well as indication of biopsy added 102 cases (20 misclassified, TN 4). Finally, if there was no biopsy, cirrhosis encounters (ICD-9 codes 571.2 and 571.5) added the last group of 110 HCV cases (26 misclassified, TN 3).

In summary, a three-criteria combination in the CART model for HBV had PPV of 87.2 %, which was significantly improved compared with 78.5 % using the six individual HBV criteria in parallel (p <0.001, n = 12,144; Table 3); while using a four-variable combination CART model for HCV identification, PPV was significantly improved to 95.2 %, compared with 83.0 % using the three initial individual HCV criteria in parallel (p <0.001)

Discussion

CHeCS is the first US longitudinal cohort study to characterize a population of over 12,000 chronic viral hepatitis patients through automated EHR and chart abstraction data collection [16]. The comprehensive data include general population characteristics, health conditions, disease-specific procedures and treatment, disease progression, and patient status. All data are derived from routine care in large healthcare systems. This EHR data platform is ideal

for Comparative Effectiveness Research; such research focuses on patient outcomes using data collected during the course of routine care to maximize clinical utility [19] and improve quality of care [8].

The HMORN VDW data have been used for collaborative research since 2003; however, it employs a distributed data model rather than a centralized database. To ensure data integrity, the CHeCS data coordinating center (DCC) centralized this data and included additional patient data collected via chart abstraction (for example, biopsy results and treatment responses). CHeCS modified the data structure by: (1) extending the VDW database CRFs by adding three EHR "VDW" structure-like CRFs; (2) adding eleven chart abstraction CRFs for comprehensive data collection; and (3) centralizing the VDW data model for stricter data control. With centralized data, the DCC can conduct efficient data processing and systematic data-quality checks.

The CHeCS data-collection process involves not only electronic data retrieval but also medical chart abstraction and patient self-reported outcomes. The electronic retrieval and medical chart abstraction datasets are complementary and provide an opportunity to assess the quality of automatic EHR data. The CART model for cohort identification is our first attempt to take advantage of such comprehensive data collection.

While it is relatively straightforward to diagnose chronic HBV/HCV in a clinical setting using serologic markers, identifying a cohort of patients with chronic HBV/HCV based on observational EHR data remains challenging in two respects. First, serologic markers may be available to a provider in a clinical setting, but may not be available in structured, queriable format (e.g., they are embedded in physician notes or external laboratory results in PDF format), or may not be complete in a general routine care setting. Second, data-processing efficiency is important when sifting through laboratory and diagnostic data for thousands of patients; therefore, e-inclusion criteria cannot be overly complex. In order to identify as many potential chronic HBV/HCV patients as possible, our original cohort identification e-criteria included a combination of laboratory results and diagnostic codes suggestive of chronic HBV or HCV; however, this "wide net" resulted in a high false-positive rate. Although we had sufficient resources to confirm case status through chart review for our initial CHeCS cohort, conducting chart reviews for 100 % of subsequent cohorts was infeasible. Furthermore, most observational studies would benefit from refining and improving electronic cohort identification criteria.

Our adaptive approach has shown that combinations of automatic e-data variables can identify accurate cohorts. Our CART models for automated EHR-based cohort identification demonstrated that a set of variable combinations could significantly improve the PPV of our original cohort inclusion criteria. The CART model for each disease was built using half of the entire cohort (thus controls for each disease consisted of the patients who met inclusion criteria for the disease and were later excluded, as well as those that met inclusion criteria for the other disease), which approximates a real clinical setting with a mixed patient population. Both CART models have been validated using the other half of the entire CHeCS cohort, and the results are robust. Each year, new patients will be added to the CHeCS cohort using the original inclusion criteria, and future chart abstraction resources

can now be directed toward those patients meeting the CART-generated criteria—improving cost-efficiency.

A potential limitation of defining cases by the presence of multiple diagnosis or procedure codes is that selection may favor patients with more severe disease, and possibly excludes cases with mild disease (and thus few or no related conditions or treatments). However, an unavoidable limitation of any cohort study based on patients engaged in medical care may be the inclusion of persons who are more ill than those in the general community. Accordingly, future analyses will be adjusted not only for demographic differences, but also for stage of disease if degree of morbidity is likely to affect results. Our adaptive CART model for patient selection is also limited to this mixed viral hepatitis population; although such a mixed population is likely to be more complex than the usual healthcare population, our model must still be validated for use in a general health system population.

In summary, CHeCS has applied several key comparative effectiveness research (CER) principles: comprehensive data collection through review of routine care records and strict data quality control; adaptive approaches to improve electronic cohort identification; and improvement to the quality of and access to hepatology care. Results from this study have already demonstrated that real-world chronic HBV patients had reduced development of hepatocellular carcinoma if they were treated with antiviral therapy [20]. The first cycle of data collection with clinical confirmation of HBV and HCV has refined the inclusion criteria used for cohort identification using CART modeling. Our adaptive approach to using electronic data for prediction of HBV and HCV infection status is feasible, can be used for sequential CHeCS cohort identification, and may be useful in other studies or clinical programs to identify patients diagnosed with HBV and HCV infection [12, 21].

Acknowledgments

CHeCS is funded by the CDC Foundation, which currently receives grants from AbbVie, Janssen Pharmaceuticals, Inc., and Vertex Pharmaceuticals. Past funders include Genentech, A Member of the Roche Group. Current and past partial funders include Gilead Sciences and Bristol-Myers Squibb. Granting corporations do not have access to CHeCS data and do not contribute to data analysis or writing of manuscripts.

Abbreviations

CHeCS	Chronic Hepatitis Cohort Study
HBV	Hepatitis B virus
HCV	Hepatitis C virus
EHR	Electronic health record
VDW	Virtual Database Warehouse
DCC	Data coordinating center
PPV	Positive predictive value
CART	Classification and Regression Trees

ROC	Receiver operating characteristic AUROC Area under the ROC curve
TN	Terminal node

References

1. Dreyer NA, Schneeweiss S, McNeil BJ, et al. GRACE principles: recognizing high-quality observational studies of comparative effectiveness. Am J Manag Care. 2010; 16:467–471. [PubMed: 20560690]
2. Alford L, AppSc B. On differences between explanatory and pragmatic clinical trials. N Z J Physiother. 2006; 35:12–16.
3. Tunis SR, Benner J, McClellan M. Comparative effectiveness research: policy context, methods development and research infrastructure. Stat Med. 2010; 29:1963–1976. [PubMed: 20564311]
4. Hastie, T., Tibshirani, R., Friedman, J. The elements of statistical learning: data mining inference and prediction. New York: Springer; 2009.
5. Johnson ML, Chitnis AS. Comparative effectiveness research: guidelines for good practices are just the beginning. Expert Rev Pharmacoecon Outcomes Res. 2011; 11:51–57. [PubMed: 21351858]
6. Iglehart JK. Prioritizing comparative-effectiveness research–IOM recommendations. N Engl J Med. 2009; 361:325–328. [PubMed: 19567828]
7. Mushlin AI, Ghomrawi H. Health care reform and the need for comparative-effectiveness research. N Engl J Med. 2010; 362:e6. [PubMed: 20054035]
8. Rongey C, Yee HF Jr. From the bedside to the community: comparative effectiveness, health services, and implementation research. Hepatology. 2011; 53:673–677. [PubMed: 21274887]
9. Moorman AC, Gordon SC, Rupp LB, et al. Baseline characteristics and mortality among people in care for chronic viral hepatitis: the chronic hepatitis cohort study. Clin Infect Dis. 2013; 56:40–50. [PubMed: 22990852]
10. Newton, KM., Larson, EB. Clin Med Res; Learning health care systems: leading through research: the 18th annual HMO research network conference; April 29–May 2, 2012; Seattle, Washington. 2012. p. 140-142.
11. HMO Resarch Network. [Accessed 3 July 2014] Figure: VDW Data Structure. Available via http://www.hmoresearchnetwork.org/asset/d53eea05-a8b6-4ea1-b4d2-6fff5e133deH/HMORN_VDW_Data_Structures.jpg
12. Spradling PR, Rupp L, Moorman AC, et al. Hepatitis B and C virus infection among 1.2 million persons with access to care: factors associated with testing and infection prevalence. Clin Infect Dis. 2012; 55:1047–1055. [PubMed: 22875876]
13. SAS Institute. SAS/STAT 9.2 user guide. Cary, NC: SAS Institution; 2010.
14. CART 6.0 user's guide salford systems [computer program] 2010.
15. Breiman, L., Friedman, J., Stone Charles, J. Classification and Regression Trees. 1. New York: Chapman and Hall; 1984.
16. Moore CL, Lu M, Cheema F, et al. Prediction of failure in vancomycin-treated methicillin-resistant Staphylococcus aureus bloodstream infection: a clinically useful risk stratification tool. Antimicrob Agents Chemother. 2011; 55:4581–4588. [PubMed: 21825294]
17. Zeger S, Liang K. Longitudinal data analysis for discrete and continuous outcomes. Biometrics. 1986; 42:121–130. [PubMed: 3719049]
18. Liang K, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986; 72:13–22.
19. Umscheid CA. Maximizing the clinical utility of comparative effectiveness research. Clin Pharmacol Ther. 2010; 88:876–879. [PubMed: 20962776]
20. Gordon SC, Lamerato LE, Rupp LB, et al. Antiviral therapy for chronic hepatitis B virus infection and development of hepatocellular carcinoma in a US population. Clin Gastroenterol Hepatol. 2014; 12:885–893. [PubMed: 24107395]
21. Holmberg SD, Spradling PR, Moorman AC, et al. Hepatitis C in the United States. N Engl J Med. 2013; 368:1859–1861. [PubMed: 23675657]

Appendix 1: CHeCS Investigators and Sites

Division of Viral Hepatitis, National Centers for HIV, Viral Hepatitis, STD, and TB Prevention (NCHHSTP), Centers for Disease Control and Prevention (CDC), Atlanta, Georgia:

Scott D. Holmberg, Eyasu H. Teshale, Philip R. Spradling, Anne C. Moorman, Jim Xing, and Cindy Tong.

Henry Ford Health System, Detroit, Michigan:

Stuart C. Gordon, David R. Nerenz, Mei Lu, Lois Lamerato, Loralee B. Rupp, Nonna Akkerman, Nancy J. Oja-Tebbe, Talan Zhang, Alexandra Sitarik, Yan Wang, and Dana Larkin.

Center for Health Research, Geisinger Clinic, Danville, Pennsylvania:

Joseph A. Boscarino, Zahra S. Daar, Robert E. Smith, and Patrick J. Curry.

The Center for Health Research, Kaiser Permanente Hawaii, Honolulu, Hawaii:

Vinutha Vijayadeva, Cynthia C. Nakasato, and John V. Parker.

The Center for Health Research, Kaiser Permanente Northwest, Portland, OR.

Mark A. Schmidt, Judy Donald, and Erin M. Keast.

Appendix 2

Chronic hepatitis B (HBV) cohort inclusion criteria:

HBV Criteria 1: Two or more ICD-9 diagnoses indicative of viral hepatitis B[1] at least 6 months apart; o

HBV Criteria 2: A viral hepatitis B or chronic liver disease ICD-9 diagnosis (571.5 cirrhosis of liver without mention of alcohol, 456.0–456.1 esophageal varices, 789.59 other ascites, 155.0 liver cancer, V42.7 liver replaced by transplant, V49.83 awaiting organ transplant status) plus positive laboratory evidence of hepatitis B surface antigen (HBsAg) or hepatitis B deoxyribonucleic acid (HBV DNA); or

HBV Criteria 3: Two or more positive laboratory results for HBsAg, hepatitis e-antigen (HBeAg), or HBV DNA, in any combination, occurring at least 6 months apart; or

HBV Criteria 4: A negative hepatitis B IgM core antibody (IgM anti-HBc) laboratory result concurrent or prior to a positive HBsAg or

[1]Chronic hepatitis B International Statistical Classification of Diseases and Related Health Problems-9th Edition (ICD-9) codes: 070.22, 070.23, 070.32, 070.33; and acute/unspecified hepatitis B ICD-9 codes: 070.20, 070.21, 070.30, and 070.31.

HBV DNA; or HBV Criteria 5: A positive hepatitis B core antibody (anti-HBc) and positive HBsAg; or

HBV Criteria 6: A positive HBsAg and an elevated alanine aminotransferase (ALT) occurring at least 6 months apart.

Chronic Hepatitis C (HCV) cohort inclusion criteria:

HCV Criteria 1: Two or more ICD-9 diagnoses indicating viral hepatitis C[1] at least 6 months apart; or

HCV Criteria 2: A viral hepatitis C[1] or qualifying chronic liver disease ICD-9 diagnosis (571.5 cirrhosis of liver without mention of alcohol, 456.0–456.1 esophageal varices, 789.59 other ascites, 155.0 liver cancer, V42.7 liver replaced by transplant, V49.83 awaiting organ transplant status) separated at least 6 months apart from a positive anti-HCV laboratory result; or

HCV Criteria 3: A positive laboratory result for hepatitis C ribonucleic acid (HCV RNA) or HCV genotype.

[1]Chronic hepatitis C diagnosis ICD-9 codes: 070.44, 070.54, 070.70, 070.71; acute/unspecified hepatitis C ICD-9 codes: 070.41, 070.51.

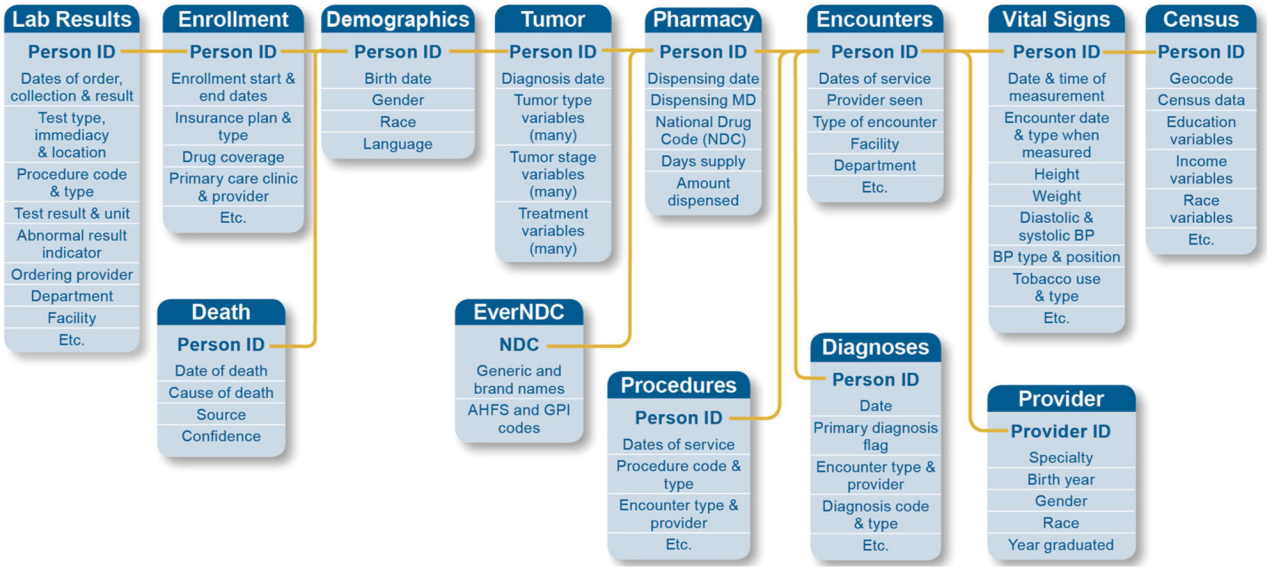
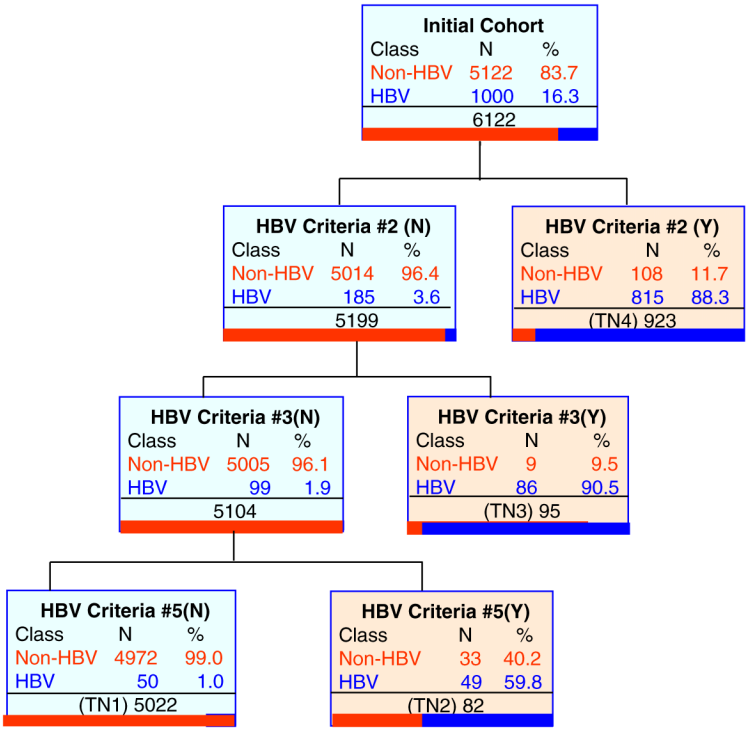


Fig. 1.
The HMORN VDW data structures [11]



HBV Criteria 2: A viral hepatitis B or chronic liver disease ICD-9 diagnosis (571.5 cirrhosis of liver without mention of alcohol, 456.0- 456.1 esophageal varices, 789.59 other ascites, 155.0 liver cancer, V42.7 liver replaced by transplant, V49.83 awaiting organ transplant status) plus positive laboratory evidence of hepatitis B surface antigen (HBsAg) or hepatitis B deoxyribonucleic acid (HBV DNA);

HBV Criteria 3: Two or more positive laboratory results for HBsAg, hepatitis e-antigen (HBeAg), or HBV DNA, in any combination, occurring at least six months apart;

HBV Criteria 5: A positive hepatitis B core antibody (anti-HBc) and positive HBsAg;

Fig. 2.
CART HBV model

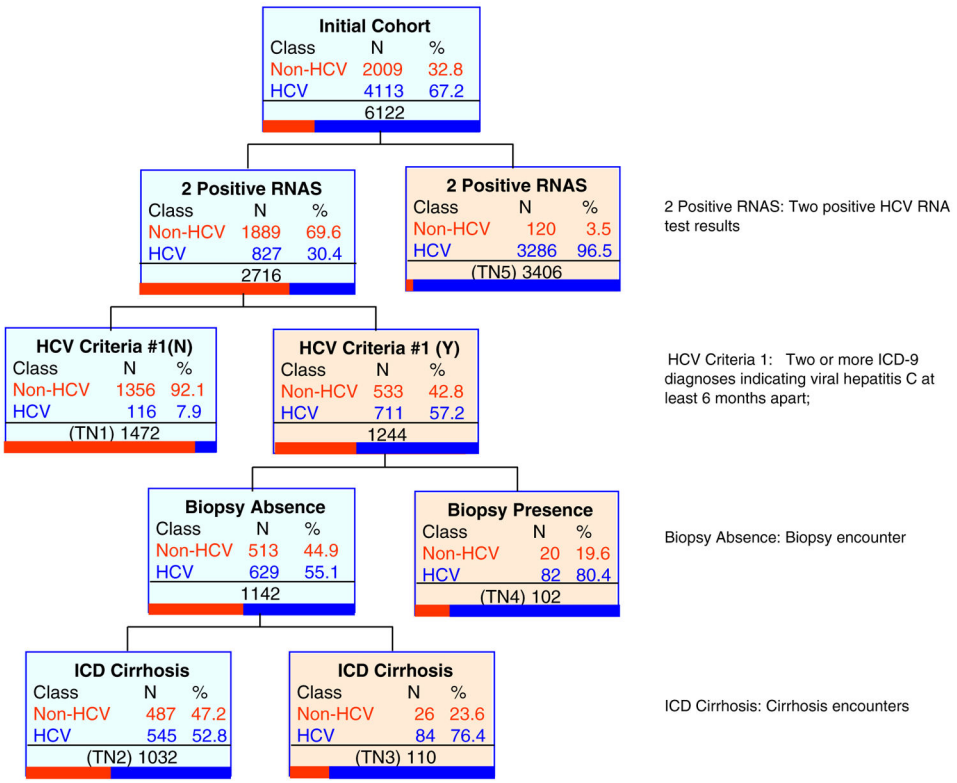


Fig. 3.
CART HCV model

Table 1

Case report form (CRF) data collection though 2008 (the initial CHeCS cohort $n = 12144$)

EHR VDW CRFs	Number of EHR records	Number of unique patients with at least one record	Abstraction/survey CRFs	Number of EHR records	Number of unique patients with at least one record
Demographics[b]	12,144	12,144	Cohort exclusion	2,992	2,956
Encounter diagnosis	2,450,034	12,123	Clinical trial drugs (hepatitis antiviral drugs only)	523	211
Encounter procedures	4,600,232	12,129	Cycle completion	31,902	11,653
Death	9,631	1,749	HBV antiviral medication starts and stops	1,405	858
Encounters	1,925,842	12,135	HCV antiviral medication change details	4,512	1,601
Census	22,197	12,067	HCV antiviral medication starts	3,621	1,605
Immunization[a]	63,918	9,473	HBV/HCV antiviral medication history	5,108	2,666
Enrollment	37,817	8,389	External labs	72,143	3,199
Payer flags[a]	36,797	12,144	Liver biopsies	6,707	4,923
Pharmacy (fills)	1,555,631	7,647	Patient's demographic updates	4,946	4,946
RX drug orders[a]	928,746	10,799	HBV resistance tests	102	93
Vital signs	340,023	10,403	Patient survey data	6,964	6,964
Laboratory results	4,769,445	12,106			
Tumor	1,416	1,266			
Total	16,753,873	134,574		140,925	41,675

[a] Additional electronic CRFs that are not VDW CRFs

[b] An extended CRF that contains HBV and HCV inclusion criteria flags as well as HBV and HCV diagnosis dates (censor time zeroes) in addition to the standard VDW variables

Table 2

CART model summary (validation results)

	HBV			HCV		
	n	**Case[b]**	**Control**	***n***	**Case[b]**	**Control**
Case[a]	992	952	40	4,058	3,430	628
Control	5,030	129	4,901	1,964	184	1,780
Total	6,022	1,081	4,941	6,022	3,614	2,408
Specificity	97.4 %			90.6 %		
Sensitivity	96.0 %			84.5 %		
PPV	88.1 %			94.9 %		
AUROC	0.97			0.93		

[a]Confirmed through chart abstraction

[b]Identified using CART

Table 3

Overall comparison

	HBV					HCV				
	n	**Original e-inclusion criteria**		**CART model**		***n***	**Original e-inclusion criteria**		**CART model**	
		Case*b*	**Control**	**Case*b***	**Control**		**Case*b***	**Control**	**Case*b***	**Control**
Case[a]	1,992	1,992	0	1,902	90	8,171	8,171	0	6,882	1,289
Control	10,152	546	9,606	279	9,873	3,937	1,680	2,293	350	3,623
Total	12,144	2,538	9,606	2,181	9,963	12,108	9,851	2,293	7,232	4,912
Specificity		94.6 %		97.3 %			58.2 %		91.2 %	
Sensitivity		100.0 %		95.5 %			100.0 %		84.2 %	
PPV		78.5 %		87.2 %			82.9 %		95.2 %	
P value[c]	< 0.0001					< 0.0001				

[a] Confirmed through chart abstraction

[b] Identified using CART

[c] GEE comparison of PPV difference between the original and CART models