BMJ Health & Care Informatics

Finding undiagnosed patients with hepatitis C virus: an application of machine learning to US ambulatory electronic medical records

John Rigg,[1] Orla Doyle,[1] Niamh McDonogh,[1] Nadea Leavitt,[2] Rehan Ali ,[1] Annie Son,[3] Bruce Kreter[3]

To cite: Rigg J, Doyle O, McDonogh N, *et al*. Finding undiagnosed patients with hepatitis C virus: an application of machine learning to US ambulatory electronic medical records. *BMJ Health Care Inform* 2023;**30**:e100651. doi:10.1136/bmjhci-2022-100651

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2022-100651).

Received 19 August 2022
Accepted 04 December 2022

[1]AI for Healthcare & MedTech, IQVIA Inc, London, UK
[2]AI for Healthcare & MedTech, IQVIA, Plymouth Meeting, Pennsylvania, USA
[3]Medical Affairs, Gilead Sciences Inc, Foster City, California, USA

Correspondence to
Dr Rehan Ali;
rehan.ali@iqvia.com

ABSTRACT

Aims To develop and validate a machine learning (ML) algorithm to identify undiagnosed hepatitis C virus (HCV) patients, in order to facilitate prioritisation of patients for targeted HCV screening.

Methods This retrospective study used ambulatory electronic medical records (EMR) from January 2015 to February 2020. A Gradient Boosting Trees algorithm was trained using patient records to predict initial HCV diagnosis and was validated on a temporally independent held-out cross-section of the data. The fold improvement in precision (proportion of patients identified by the algorithm who are HCV positive) over universal screening was examined and compared with risk-based screening.

Results 21 508 positive (HCV diagnosed) and 28.2M unlabelled (lacking evidence of HCV diagnosis) patients met the inclusion criteria for the study. After down-sampling unlabelled patients to aid the algorithm's learning process, 16.2M unlabelled patients entered the analysis. Performance of the algorithm was compared with universal screening on the held-out cross-section, which had an incidence of HCV diagnoses of 0.02%. The algorithm achieved a 101.0 ×, 18.0 × and 5.1 × fold improvement in precision over universal screening at 5%, 20% and 50% levels of recall. When compared with risk-based screening, the algorithm required fewer patients to be screened and improved precision.

Conclusions This study presents strong evidence towards the use of ML on EMR data for the prioritisation of patients for targeted HCV testing with potential to improve efficiency of resource utilisation, thereby reducing the workload for clinicians and saving healthcare costs. A prospective interventional study would allow for further validation before use in a clinical setting.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Hepatitis C virus (HCV) is one of the most common blood-borne virus and is the target of a WHO initiative to eradicate it as a public health threat by 2030. Universal one-time screening for adults has been recommended in the USA; however, this is challenging to implement in practice, and screening rates remain low.

⇒ Machine learning approaches for finding undiagnosed HCV patients have been favourably evaluated using retrospective health claims data in the past, introducing the potential for more targeted and effective screening programmes.

WHAT THIS STUDY ADDS

⇒ This study develops machine learning methods to predict potentially undiagnosed HCV patients using a large-scale, retrospective, US, ambulatory electronic medical record (EMR) data set.

⇒ It adds to current knowledge since analysis is based on choice of a more appropriate data set which, critically, corresponds to the setting in which an algorithm would be implemented. Moreover, various methodological choices (such as a temporally separate held-out set for model evaluation) lead to greater clinical insight and more robust predictions than elsewhere in the literature.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This study suggests that machine learning algorithms, if integrated into EMR systems and clinical workflows, would enable targeted HCV screening, thus accelerating progress towards HCV elimination.

INTRODUCTION

Hepatitis C virus (HCV) is one of the most common blood-borne viruses and a major cause of liver-related morbidity and mortality in the USA.[1] The estimated prevalence of HCV in the USA is 1%[2] with the estimated number of new (acute) infections increasing fourfold between 2010 and 2018.[3] Treatment of HCV has been revolutionised in recent years by direct-acting antiviral drugs which are well tolerated and highly efficacious (>95% cure rate).[4–6] These developments paved the way for the WHO to propose a global strategy to eliminate HCV as a public health threat by 2030.[7] In the USA, the National Academies of Science, Engineering and Medicine developed an HCV elimination plan where improved detection of undiagnosed cases is a

key element.[8] This, together with the need for identifying hard-to-find patients not captured by risk-based screening, has led to increased emphasis on universal one-time HCV screening recommended as part of the American Association for the Study of Liver Diseases (AASLD) - Infectious Diseases Society of America (IDSA) guidance as well as periodic screening in high-risk individuals.[6] Recent studies show that HCV screening rates remain low and recommend targeted interventions aimed at patients and physicians to boost screening rates.[9 10]

The advent of electronic medical records (EMR) used in combination with machine learning (ML) has presented new opportunities for screening in population health management.[11 12] EMRs have been used previously to find undiagnosed HCV cases;[13–15] however, these studies use simple clinical rules to prioritise patients for HCV screening. Previous work has demonstrated how ML can accurately identify undiagnosed HCV cases using US medical insurance claims and prescription data.[16] Additionally, ML techniques applied to EMRs have been used for patient finding in other disease areas, such as type 1 diabetes and sepsis.[17 18] Given the promise shown in applying ML to EMRs, we investigated whether undiagnosed HCV cases could be predicted by an ML algorithm using a US EMR data set. The Methods section describes how this was developed and, in the results, a benchmark of performance against universal and risk-based screening is provided. Finally, the discussion contains an appraisal of how prioritisation of patients in the US for HCV screening could be improved with the algorithm, along with the potential impact on resource utilisation and the subsequent prospective validation requirements.

METHODS

Study design

This retrospective, observational study used anonymised medical records between January 2015 and February 2020 from the IQVIA Ambulatory Electronic Medical Records (AEMR) database covering over 80M US patients.

Patient selection

The algorithm developed in this study predicts HCV patients, including undiagnosed current infections and new cases over the next year (which are detected in the clinic by HCV antibody and/or RNA tests). The algorithm was trained on patients aged 12 years and over with evidence of healthcare utilisation during their lookback period, who were assigned to either a positive or unlabelled cohort. The positive cohort was defined as patients who have a diagnosis code (including for acute, chronic, carrier and unspecified HCV types) or treatment code relating to their first HCV record over a 12-month selection window (online supplemental tables S1, S2). Patients with HCV records outside of this selection window were excluded. The unlabelled cohort was defined as patients with no evidence of HCV infection throughout their medical history, which likely includes HCV-positive patients missing a formal diagnosis label. (This makes it representative of the population that the model will be applied to in real-world use.) The unlabelled cohort was down sampled to reduce the effect of class imbalance on algorithm development.[19] For validation, all results were projected to the expected number of unlabelled patients in the deployment setting, that is, the count of false positives was projected to match the expected number of non-HCV patients in the selected population.

Predictor selection

For predictor selection, stakeholders with clinical expert knowledge were invited to define events relevant to HCV, which spanned diagnoses, prescriptions, procedures and lab tests, and comprised of 276 predictors, including demographics (online supplemental table S3). These predictors were mapped to clinical codes by coding experts and extracted over the lookback period. These predictors were described by their frequency and timing (recency, duration, initial onset); in the case of lab test results, the earliest and most recent values, delta, average, maximum and minimum values were also captured. This resulted in a total of 1175 predictors. Predictors that were present in less than 0.1% of the positive and unlabelled cohorts were removed.

The predictor that captured risk of substance abuse is referred to as Risk of being a Person Who Injects Drugs (R-PWID) and was subsequently used to benchmark performance. It was defined as an International Classification of Diseases (ICD) claim for substance abuse and/or withdrawal, prescription for substance abuse agents.

Machine learning algorithm

An ML algorithm was developed to learn the prediagnosis journey of HCV patients, which can then be applied to novel patient data to compute a risk score for HCV ranging between 0 and 1 (online supplemental figure S1). Gradient Boosting Trees (GBTs)[20] were chosen as the ML algorithm due to their ability to handle missing data, sensitivity to interactions and non-linear relationships, approaches for controlling overfitting, robustness to noisy/mislabelled data (including the presence of undiagnosed positives in the unlabelled class) and success in structured healthcare data problems.[16 21–23] Additionally, GBTs are compatible with the model explanation technique Shapley Additive exPlanations (SHAP) (see “Interpretation” section in Methods).[24]

Implementation and validation

The GBT algorithm was trained and tested using cross-sections with non-overlapping selection windows (online supplemental figure S2), with the most recent cross-section held out as an independent test set. The study was devised using a rolling cross-sectional design, whereby a 24-month lookback period is followed by a 12-month selection window (see online supplemental figure S1). The lookback period describes the prediagnosis medical history seen by the ML algorithm, and the selection

window is used to define the subsequent outcome being predicted. The trade-off between algorithm complexity (number of predictors) and performance was analysed. A full description of the pipeline is found in online supplemental information (S1.1). Given the imbalanced nature of the task (ie, many more unlabelled than positive patients are expected) precision (positive predictive value; proportion of patients identified by the ML algorithm that are HCV positive) and recall (sensitivity; proportion of HCV cases identified by the ML algorithm) were selected to assess model performance.[25]

To provide context, precision is benchmarked against risk-based screening approaches R-PWID and the 1945–1965 'Baby Boomer' birth cohort. Here, precision is calculated as the number of patients in the HCV cohort who meet the risk-based criteria divided by the total number of patients who meet these criteria. Precision is reported at recall levels corresponding to the proportion of undiagnosed HCV cases identified by the risk-based approaches.

The fold improvement in precision over universal screening on the held-out cross-section is reported. Additionally, the scaled precision (ie, projected to the incidence of HCV in the data) and specificity at 5%, 20% and 50% recall are reported, along with the area under the receiver operating characteristics curve (AUROC), where a value of 1 indicates that the model can perfectly separate the two classes, and 0.5 indicates that the performance is equivalent to random chance.

Interpretation

The interpretation of the ML algorithm will focus on the importance and dynamics of the predictors which will be described at the global level (across all patients) and for local patient subgroups. This is facilitated by the SHAP methodology, which quantifies how each predictor contributes to the risk score for an individual patient.[26] This method accounts for key limitations in classical predictor importance estimation, such as correlation between variables, by considering all possible sets and orderings of features in a computationally efficient manner.[27]

Testing for algorithmic bias

The ML risk scores were tested for unintended algorithmic bias across the protected characteristics: age, gender and race. Given the intended use, a false negative would result in a patient being deprioritised for screening, which would have a higher consequence than a false positive. Therefore, equal opportunity was tested by comparing false-negative rates across characteristic subgroups.[28] A post hoc approach for univariate correction of algorithmic bias was applied by calculating thresholds for corresponding recall levels within each protected characteristic subgroup.[28] In practice, this translates to screening a larger number of patients belonging to the subgroups that bias is operating against.

RESULTS

Characteristics of study population

For the positive and unlabelled cohorts, 21 508 and 28.2M patients met the selection criteria, respectively. The incidence of HCV in the held-out cross-section was 0.02%, which corresponds to the precision for universal screening. The unlabelled cohort was down sampled at a cross-section level to reduce the imbalance between cohorts, resulting in 16.2M non-HCV patients entering the analysis. After excluding predictors with ultra-low prevalence, 931 out of 1175 predictors were retained. The patient demographics and key risk factors are summarised in Table 1 for the patients in the held-out cross-section. Highly similar patient characteristics are observed for all patients versus the test cross-section. As expected, there are higher rates of R-PWID, opioid use, HIV infection and cirrhosis in the positive cohort as well as higher rates of chronic disorders, such as psychiatric disorders and diabetes.

Figure 1 illustrates the recency of clinical events with respect to first HCV diagnosis. Opioid and non-opioid analgesics were observed in 30% and 43% of HCV patients, respectively, within the 5 months prior to diagnosis. Substance dependence was observed through prescription of relevant agents (8%), diagnoses for substance abuse (15%) and withdrawal (4%) and occurred most recently an average of 4 months prior to diagnosis. The most common specialty visited was Family Practice.

Model performance

The universal screening approach screens the full patient population and so would identify all undiagnosed HCV cases. However, as this has a high burden on healthcare providers, risk-based screening is used to screen fewer, high-risk individuals. In the held-out cross-section, R-PWID screening finds 20.9% of HCV cases (from screening, 3.7% of the population) and the 1946–1964 birth cohort finds 48.4% of HCV cases (from screening 34.5% of the population). Conversely, the proposed ML algorithm provides a unique solution, whereby either the proportion of HCV cases identified or the proportion of the patient population tested can be predefined. Therefore, to compare the algorithm's performance against the risk-based approaches, we can evaluate its precision at the same recall levels as the ones achieved by the risk-based methods, that is, at the same proportion of identified HCV cases (see table 2). Table 2 shows how fewer patients need to be screened when the algorithm is used to prioritise patients.

At 5%, 20% and 50% recall, the algorithm's precision was 2%, 0.4% and 0.12%, and specificity was 99.9%, 99.0% and 90%, respectively. The AUCROC was 0.81. The algorithm's precision can be compared with universal screening, if universal screening is adapted such that patients are randomly selected from the population for screening until either 5%, 20% or 50% of undiagnosed HCV cases are identified. We assume that the precision of screening these proportions is equivalent to universal

Table 1 Patient characteristics over the cross-sectional lookback period.

	All (unique patients)		Held-out cross-section	
	HCV	Non-HCV*	HCV	Non-HCV*
Patient counts (N)	21 508	28 215 073	4641	21 246 498
Age (years; mean (±SD))	52±15.2	52±20.0	53±15.6	53±20.4
Gender (% male)	51.4%	42.0%	50.5%	41.9%
HCV relevant predictors				
R-PWID (%)	20.6%	4.8%	20.9%	3.7%
1946–1964 birth cohort (%)	51.0%	34.0%	48.4%	34.5%
Opioid usage (%)	36.1%	21.4%	33.3%	19.4%
Cirrhosis (%)	1.9%	0.2%	1.9%	0.3%
HIV/Aids (%)	1.7%	0.5%	1.6%	0.5%
Chronic liver disease (%)	5.7%	1.8%	4.8%	1.9%
Chronic disorders				
Anxiety (%)	19.5%	12.1%	18.5%	11.7%
Chronic lung disease (%)	15.4%	10.4%	14.4%	10.0%
CKD or ESRD (%)	4.9%	4.0%	5.0%	4.1%
Depression (%)	18.3%	11.1%	17.1%	10.6%
Diabetes (%)	16.5%	13.7%	16.7%	13.5%
Hyperlipidaemia (%)	22.2%	26.2%	21.3%	25.8%
Hypertension (%)	32.3%	27.5%	32.9%	27.0%

*Counts projected to account for down-sampling.
CKD, chronic kidney disease; ESRD, end stage renal disease; HCV, hepatitis C virus; R-PWID, risk of being a person who Injects drugs.

screening (12-month incidence in the held-out data, 0.02%). Here, the algorithm has 101.0×, 18.0× and 5.1× fold improvement in precision over universal screening at 5%, 20% and 50% recall. The performance versus complexity analysis revealed the number of predictors can be reduced from 1175 to 100 without negatively impacting performance (online supplemental figure S3). Further reductions in the number of predictors are feasible with no observable impact on performance at high recall levels (ie, >20%).

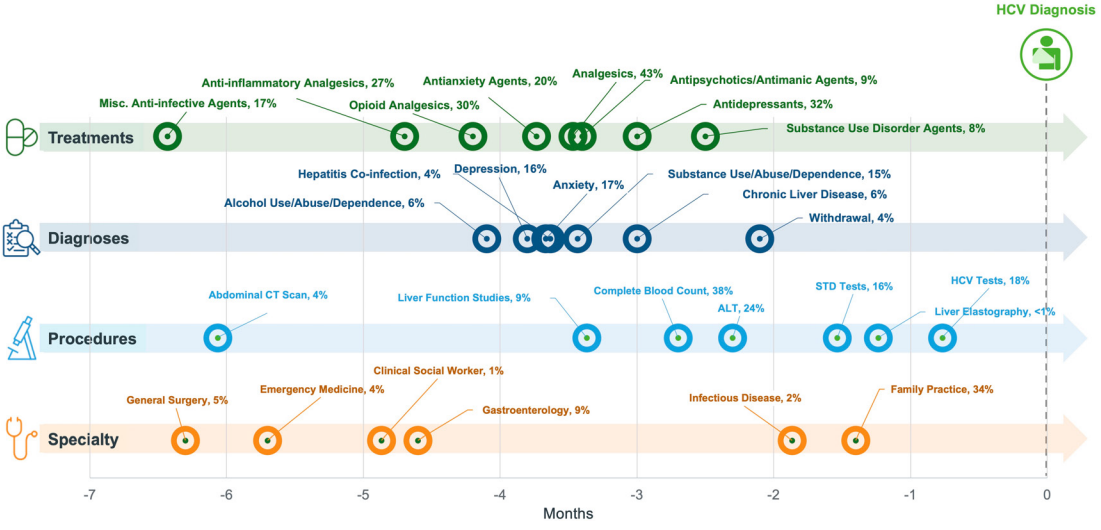


Figure 1 Patient journey in the months leading up to their first observed diagnosis of HCV. The graphic displays median time prior to HCV diagnosis of the most recent event with the most prevalent events chosen for illustration. Note that patient characteristics are plotted with respect to the timing of the first exposure to HCV. In contrast, the patient characteristics in table 1 are with respect to the beginning of the lookback period which is provided to the ML algorithm, that is, patient history which occurs during the selection window is included in figure 1 but excluded in table 1. HCV, hepatitis C virus; ML, machine learning.

Table 2 Performance of ML algorithm compared with risk-based screening

Approach	Proportion of the overall patient population to be screened by each approach	Proportion of correctly identified undiagnosed HCV cases (precision)
To find 20.9% of undiagnosed HCV cases		
R-PWID cohort	3.70%	0.12%
ML algorithm	1.80%	0.40%
To find 48.4% of undiagnosed HCV cases		
Birth cohort	34.50%	0.03%
ML algorithm	12.30%	0.12%

HCV, hepatitis C virus; ML, machine learning; R-PWID, risk of being a person who injects drugs.

Interpretation

The contribution of predictors to the ML risk score for HCV is displayed in figure 2 using two views: (1) contributions averaged across all patients and (2) the 100 highest scoring patients and the 100 lowest scoring patients. From the global view, patient demographics and age play an important role in the prediction of HCV as well as the use of analgesics (both opioid and non-opioid), hyperlipidaemia and lab test results for aspartate transaminase (AST). The local view for the highest scoring patients reveals strong contributions from predictors capturing substance abuse, the number of HCV tests in recent history and AST lab results, with age playing a minor role. In contrast, for the lowest scoring patients, age plays a dominant role in determining their risk score with minor contributions from lab test results, use of non-opioid analgesics, race and hypertension. In online supplemental figure S4, the interaction between age and gender is described by plotting the contribution of age to a patient's risk score and grouping by gender. For age, we see a bimodal dynamic, with patients between 25 and 35 years and patients between 50 and 70 years having higher risk of HCV. In particular, women in the first age bracket are assigned a higher risk score than men, with the reverse observed in the second age bracket. Note that this is not a causal analysis and the associations may be driven by other factors, such as pregnancy in women enabling more regular touchpoints with the healthcare system or proactive screening for HCV.

Algorithmic bias

The false-negative rates for each protected characteristic across the 5% incremental recall bins are shown in figure 3. For age, the false-negative rates are highest for patients aged 75 and over; for gender, they are marginally higher for women than men; for race, they are highest for Asian, Hispanic, other and unknown, indicating algorithmic bias against these subgroups. Post hoc correction can ensure equal opportunity for a single characteristic but not multiple characteristics in combination, as shown in online supplemental figures S5-S7.

DISCUSSION

The ML algorithm showed an increased efficiency of screening for HCV compared with universal screening and risk-based approaches, where fewer patients are

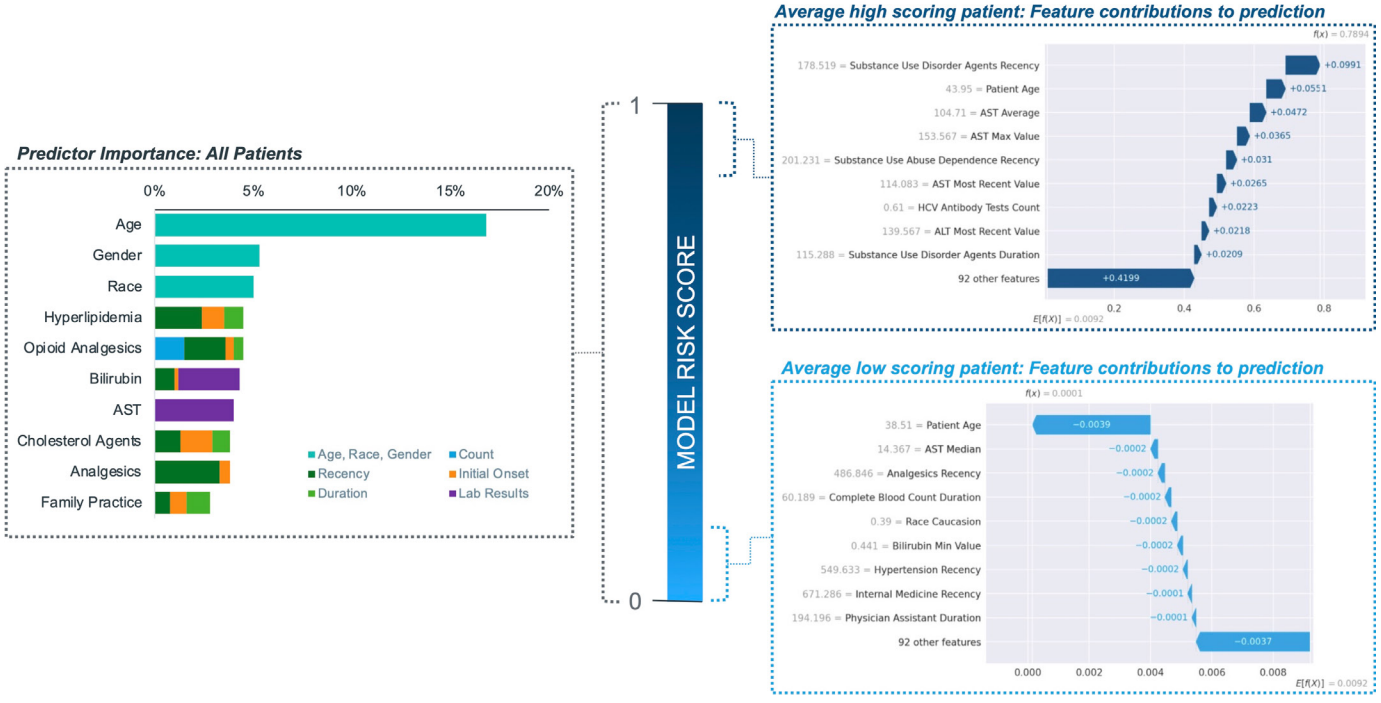


Figure 2 Predictor importance globally (L) and for most extreme patients (R). AST, aspartate transaminase; HCV, hepatitis C virus; ALT, alanine aminotransferase.

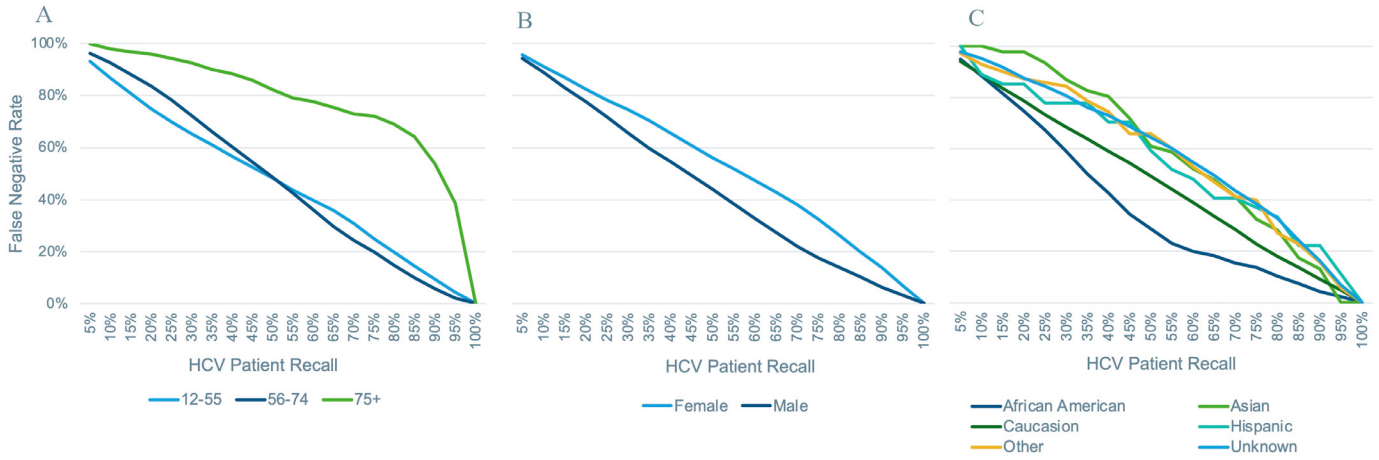


Figure 3 Subgroup false-negative rate versus per HCV patient recall (across all subgroups) by the protected characteristics; (A) age, (B) gender and (C) race. HCV, hepatitis C virus.

required to be screened with the algorithm to identify the equivalent number of HCV cases. This supports existing research that found ML algorithms trained on EMR data can be used to predict patients' risk of disease with high precision.[16–18] Moreover, this study demonstrates the utility of an EMR-based ML algorithm in identifying HCV patients and evidences a potential benefit in deployment into clinical workflow. One way to realise this benefit is through integrating risk prediction algorithms into EMR systems, and examples of this exist; simple rule-based algorithms have been effective in increasing HCV screening rates,[15] while a recent study describes the integration of a complex sepsis prediction of ML algorithm with an Epic EMR system in the USA.[17] EMR integration can facilitate targeted HCV screening, which would have multiple potential clinical and operational benefits. First, effective targeting can improve the allocation of limited healthcare resources and hence the return on investment for a screening programme. Second, effective targeting would be expected to lead to improvements in rates of HCV diagnosis, treatment and transmission as well as reductions in morbidity and mortality arising from earlier diagnosis. Third, a sophisticated risk-based targeting approach can identify hard-to-find patients who may be overlooked by simple screening criteria. Finally, the algorithm outputs a continuous risk score enabling a nuanced triage process. For instance, patients with high risk scores could be proactively invited for screening, whereas patients with lower risk scores could be opportunistically screened during routine visits.

There is a need to understand biases in ML models. The ML algorithm developed here exhibits signs of representation bias (which arises through lack of generalisation to groups that are under-represented in the data). A post hoc univariate corrective approach showed promise in reducing bias across a single characteristic. This approach calculates how many patients from each characteristic's subgroups should be screened to equalise the proportion of HCV patients identified belonging to each. However, when a single characteristic is equalised with this approach, it may worsen bias for others. Therefore, a more expansive approach to address all characteristics equitably would form part of future work.

The scope of this study is restricted to individuals who have engaged with the US healthcare system. In a future deployment setting, this would result in low chance of prioritisation for people with limited or no access to healthcare in the USA. This is particularly relevant for HCV as a high proportion of individuals infected with HCV is either uninsured or have publicly funded health insurance.[29] Therefore, complementary approaches are needed, such as routine HCV screening in addiction medicine settings, correctional facilities and proactive HCV screening in sexual health settings, alongside investment in HCV treatment networks to ensure linkage to care is facilitated.[30 31]

The results of this study represent a proof of concept that has been developed using a US-based EMR data set. A natural next step for this algorithm is to perform further validation in an interventional prospective study that emulates the real-world deployment settings. This will help overcome some limitations of the retrospective study design. In particular, the positive cohort in this study comprises of patients who are diagnosed over a finite outcome window in the absence of the intervention of interest (ie, screening of the identified patients). Therefore, the number of false positives is overestimated for each screening intervention (universal, ML-algorithm, etc).

An important additional dimension to this study is the cost-effectiveness of the ML algorithm for screening. Previous studies have reported that risk-based HCV screening in populations such as PWID and the Baby Boomer birth cohort are cost-effective.[30 31] Given the ML algorithm has further increased efficiency, it is likely that this will translate into a further increase in cost-effectiveness. A formal study of the cost-effectiveness of the ML algorithm will form an important part of future work.

This study presents strong evidence to support the use of an HCV prediction ML algorithm with large-scale EMR data. The focus of the work will now move to a pilot phase involving integration and prospective interventional validation of the algorithm in a clinical research setting. Subject to a successful pilot study, focus will shift to local deployment of the algorithm in multiple healthcare settings and geographies, which will involve collaboration with end users and on-going monitoring, with the ultimate goal of contributing to efforts towards HCV elimination by targeted increase in diagnosis rates and reducing time to diagnosis.

Contributors JR planned the study, provided methodology support, reviewed the manuscript and is guarantor. OD managed the study and supervised the work by NMD. NMD executed the technical work and wrote the first draft of the manuscript. NL planned the study, provided methodology support and reviewed the manuscript. RA reviewed the technical work by NMD, wrote the second draft of the manuscript and submitted the study. AS planned the study and served as project manager for the study. BK provided scientific leadership for the study.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available. All data belongs to IQVIA.

ORCID iD

Rehan Ali http://orcid.org/0000-0003-3182-2106

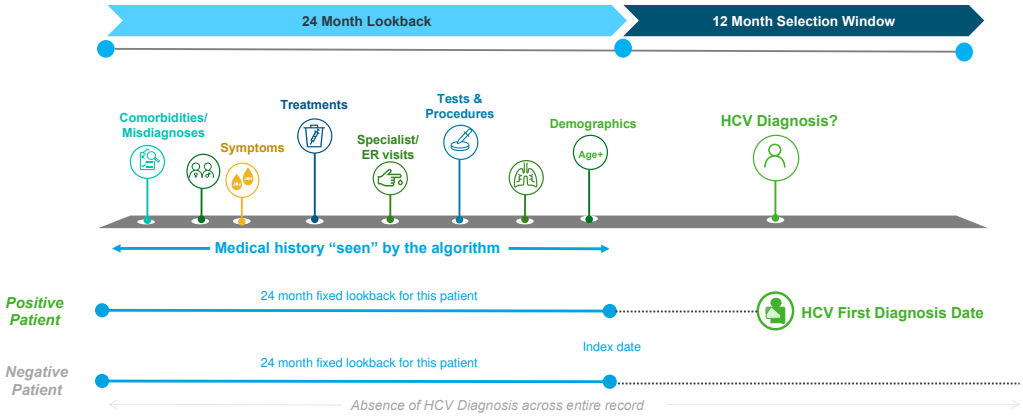
REFERENCES

1 Asrani SK, Devarbhavi H, Eaton J, *et al*. Burden of liver diseases in the world. *J Hepatol* 2019;70:151–71.
2 Hofmeister MG, Rosenthal EM, Barker LK, *et al*. Estimating prevalence of hepatitis C virus infection in the United States, 2013-2016. *Hepatology* 2019;69:1020–31.
3 Prevention CfDCa. *Viral hepatitis statistics and Surveillance—United states*, 2018.
4 Falade-Nwulia O, Suarez-Cuervo C, Nelson DR, *et al*. Oral direct-acting agent therapy for hepatitis C virus infection: a systematic review. *Ann Intern Med* 2017;166:637–48.
5 European Association for the Study of the Liver. Electronic address eee, Clinical Practice Guidelines Panel C, representative EGB, Panel m. EASL recommendations on treatment of hepatitis C: Final update of the series(). *J Hepatol* 2020;73:1170–218.
6 Ghany MG, Morgan TR, AASLD-IDSA Hepatitis C Guidance Panel. Hepatitis C guidance 2019 update: American association for the study of liver Diseases-Infectious diseases Society of America recommendations for testing, managing, and treating hepatitis C virus infection. *Hepatology* 2020;71:686–721.
7 WHO. Combating hepatitis B and C to reach elimination by 2030, 2021. Available: https://www.who.int/hepatitis/publications/hep-elimination-by-2030-brief/en/ [Accessed 09 Mar 2021].
8 NASEMStrom BL, Buckley GJ, eds. *A national strategy for the elimination of hepatitis B and C: phase two report*, 2017.
9 Kasting ML, Christy SM, Reich RR, *et al*. Hepatitis C virus screening: factors associated with test completion in a large academic health care system. *Public Health Rep* 2022;137:1136–45.
10 Kasting ML, Giuliano AR, Reich RR, *et al*. Hepatitis C virus screening trends: serial cross-sectional analysis of the National health interview survey population, 2013-2015. *Cancer Epidemiol Biomarkers Prev* 2018;27:503–13.
11 Flaxman AD, Vos T. Machine learning in population health: opportunities and threats. *PLoS Med* 2018;15:e1002702.
12 Morgenstern JD, Buajitti E, O'Neill M, *et al*. Predicting population health with machine learning: a scoping review. *BMJ Open* 2020;10:e037860.
13 Burrell CN, Sharon MJ, Davis S, *et al*. Using the electronic medical record to increase testing for HIV and hepatitis C virus in an Appalachian emergency department. *BMC Health Serv Res* 2021;21:524.
14 Zucker J, Aaron JG, Feller DJ, *et al*. Development and validation of an electronic medical record–based algorithm to identify patient milestones in the hepatitis C virus care cascade. *Open Forum Infect Dis* 2018;5:ofy153.
15 Barter L, Cooper CL. The impact of electronic medical record system implementation on HCV screening and continuum of care: a systematic review. *Ann Hepatol* 2021;24:100322.
16 Doyle OM, Leavitt N, Rigg JA. Finding undiagnosed patients with hepatitis C infection: an application of artificial intelligence to patient claims data. *Sci Rep* 2020;10:10521.
17 Sendak MP, Ratliff W, Sarro D, *et al*. Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. *JMIR Med Inform* 2020;8:e15182.
18 Cheheltani R, King N, Lee S, *et al*. Predicting misdiagnosed adult-onset type 1 diabetes using machine learning. *Diabetes Res Clin Pract* 2022;191:110029.
19 Liu X-Y, Wu J, Zhou Z-H. Exploratory undersampling for class-imbalance learning. *IEEE Trans Syst Man Cybern B Cybern* 2009;39:539–50.
20 Chen T, Guestrin C. XGBoost : Reliable Large-scale Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2016:785–94.
21 Zhang Z, Zhao Y, Canes A, *et al*. Predictive analytics with gradient boosting in clinical medicine. *Ann Transl Med* 2019;7:152.
22 Doyle OM, van der Laan R, Obradovic M, *et al*. Identification of potentially undiagnosed patients with nontuberculous mycobacterial lung disease using machine learning applied to primary care data in the UK. *Eur Respir J* 2020;56. doi:10.1183/13993003.00045-2020. [Epub ahead of print: 01 10 2020].
23 Baher HL, Lemaire V, Trinquart R. On the intrinsic robustness of noise of some leading classifiers and symmetric loss function - an empirical evaluation. *arXiv* 2010:13570 [cs.LG].
24 Lundberg S, Lee S-I. A unified approach to interpreting model predictions, 2017. Available: https://ui.adsabs.harvard.edu/abs/2017arXiv170507874L [Accessed 01 May 2017].
25 Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, 2020. Available: https://ui.adsabs.harvard.edu/abs/2020arXiv201016061P [Accessed 01 Oct 2020].
26 Lundberg S, Lee S-I. *A unified approach to interpreting model predictions. presented at: advances in neural information processing systems*, 2017.
27 Lundberg SM, Erion G, Lee S-I. Consistent individualized feature attribution for tree ensembles 2018:abs/1802.03888.
28 Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning, 2016. Available: https://ui.adsabs.harvard.edu/abs/2016arXiv161002413H [Accessed 01 Oct 2016].
29 Ong JP, Collantes R, Pitts A, *et al*. High rates of uninsured among HCV-positive individuals. *J Clin Gastroenterol* 2005;39:826–30.
30 Barbosa C, Fraser H, Hoerger TJ, *et al*. Cost-effectiveness of scaling-up HCV prevention and treatment in the United States for people who inject drugs. *Addiction* 2019;114:2267–78.
31 Coward S, Leggett L, Kaplan GG, *et al*. Cost-effectiveness of screening for hepatitis C virus: a systematic review of economic evaluations. *BMJ Open* 2016;6:e011821.

Supplementary Information for "Finding undiagnosed patients with Hepatitis C Virus: an application of artificial intelligence to US ambulatory electronic medical records"

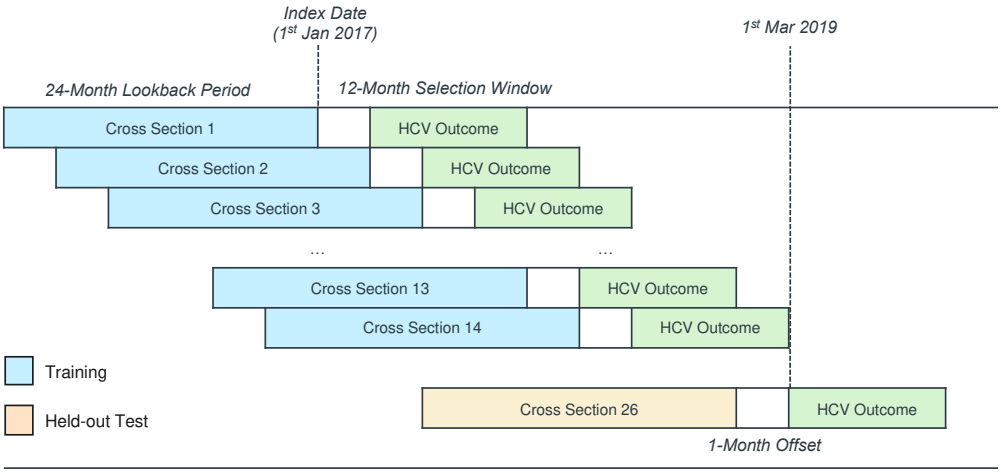
1 Supplementary Methods

Figure S 1 Study design: cross-sectional approach.



The study was designed as a retrospective, cross-sectional database study. Patients were assigned to either the HCV or non-HCV cohort as described in the main text (see also Figure S1). The diagnosis and product codes used to define HCV are listed in Tables S1 and S2. Cross-sections were extracted on a rolling basis with between January 2015 and February 2020 with the final cross-section designed to have a non-overlapping selection window to facilitate subsequent validation of the model, see Figure S2. The ML algorithm is depicted above for a single cross-section. It shows medical history "seen" by the algorithm in the 24 month look back for the patient and how the algorithm predicts a HCV diagnosis in the 12 month selection window.

Figure S 2 Rolling cross-sectional study design.



1.1 Machine Learning Algorithm: Implementation and Validation

The GBT algorithm was executed using the XGBoost (xgboost v1.2.1) implementation for Python (3.6.8). The GBT algorithm was trained using cross-sections 1 to 14 and subsequently tested on cross-section 26, where the selection window did not overlap with the training cross-section selection windows.

The non-HCV cohort was randomly down-sampled to a ratio of 100 non-HCV to HCV patients within each cross-section. This ratio was chosen to reduce the class imbalance whilst preserving the heterogeneity of the non-HCV cohort. After down-sampling, a selection criterion that requires each patient to have a predictor in the lookback period was applied. When assessing the model performance on the test cross-section, the number of non-HCV patients was rescaled to the ratio seen within the underlying population. This was to account for the artificially low number of non-HCV patients which would result in an artificially low false positive rate.

Model complexity was optimised by reducing the number of features iteratively. An initial model was trained on the full predictor space (931 predictors) using the earliest two cross-sections. This model was applied to cross-section 14, a left out and non-overlapping training cross-section, and performance was reported as improvement in precision over Universal Screening at recall levels of 5%, 10%, 20%, 50% and 75%. Subsequent models were retrained iteratively, reducing the predictor space to only the most important predictors as identified by the total gain, i.e. the contribution of splitting on the predictor to model performance. The model with the lowest number of predictors without any reduction in performance was chosen.

The training of GBT algorithm included hyperparameter tuning for the learning rate, the number of estimators, max depth, min child weight, and gamma using the grid search method in a cross-validated manner.

The final step involved training the GBT algorithm with the optimised hyperparameters on all available training data followed by its application to the held-out cross-section to assess model performance.

2 Supplementary Results

Figure S3 shows model performance as improvement over Universal Screening versus model complexity (the number of model features) at recall levels of 5%, 10%, 20%, 50% and 75%. The 100-predictor model was chosen as it reduced complexity whilst retaining model performance.

Figure S 3 Performance versus complexity (number of predictors).

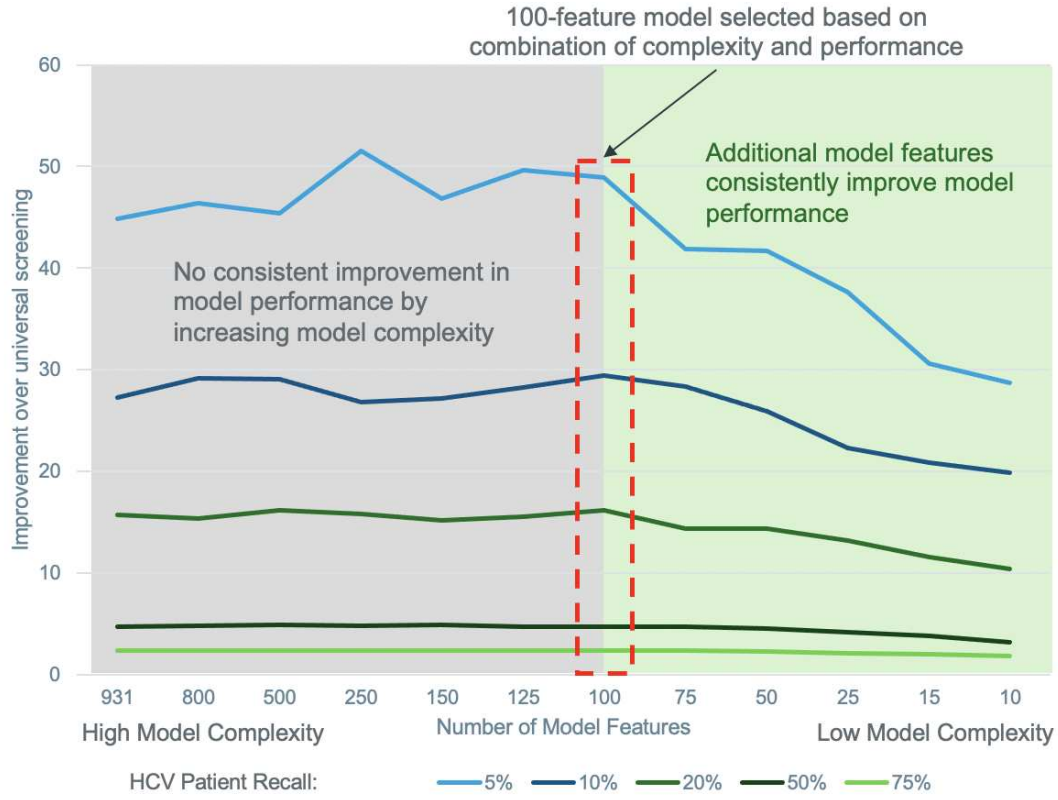


Figure S 4 Contribution of age and gender to the HCV risk score where each patient is represented to a single data point.

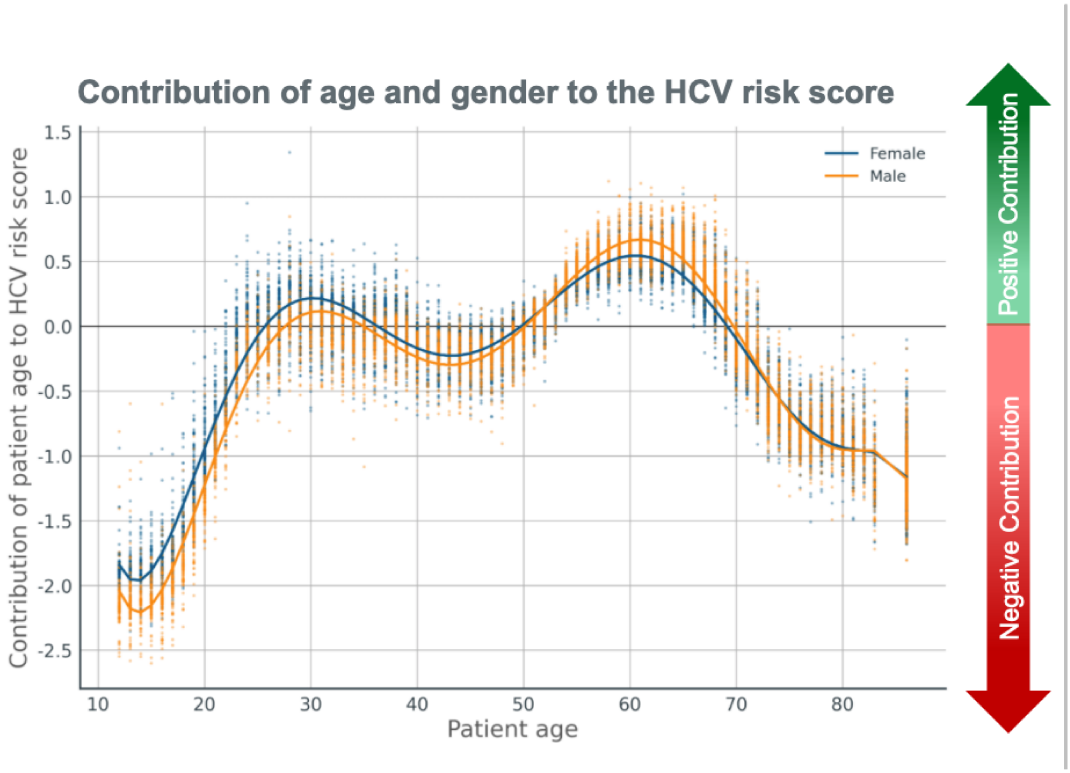


Figure S 5 False Negative Rate per HCV Patient Recall post correction for bias in Age by the protected characteristics; a) Age b) Gender and c) Race

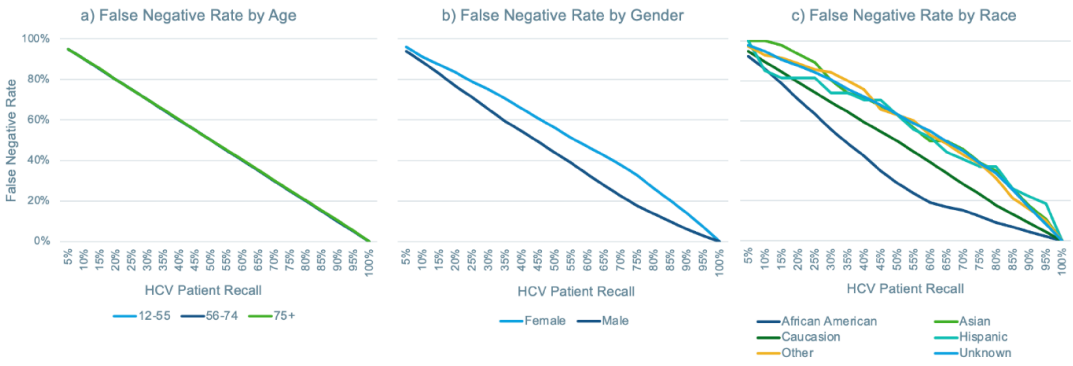


Figure S 6 False Negative Rate per HCV Patient Recall post correction for bias in Gender by the protected characteristics; a) Age b) Gender and c) Race

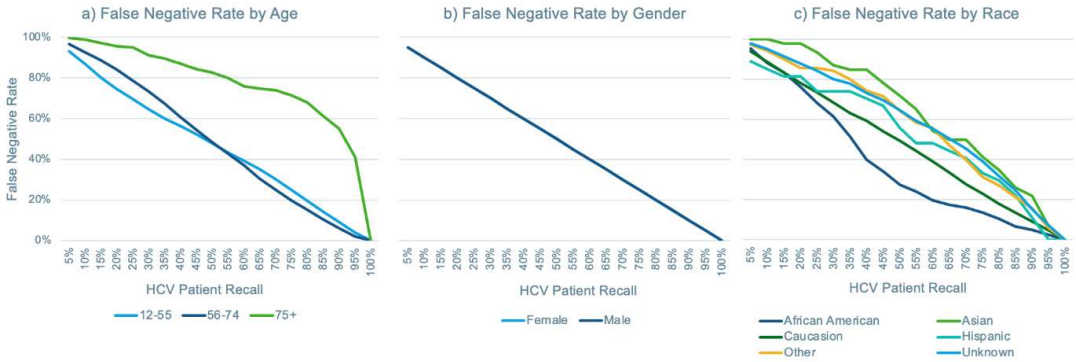
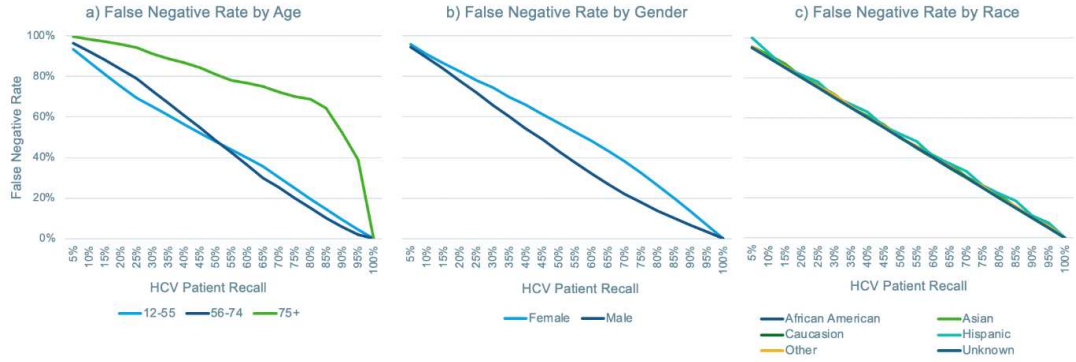


Figure S 7 False Negative Rate per HCV Patient Recall post correction for bias in Race by the protected characteristics; a) Age b) Gender and c) Race



3 Supplementary Tables

3.1 Diagnosis codes and prescription products for HCV

Table S 1 List of ICD 9 and ICD 10 codes used to select HCV patients.

DIAGNOSIS CODE TYPE	DIAGNOSIS CODE	DIAG DESCRIPTION
ICD 9	070.41	ACUTE HEPATITIS C WITH HEPATIC COMA
ICD 9	070.44	CHRONIC HEPATITIS C WITH HEPATIC COMA
ICD 9	070.51	ACUTE HEPATITIS C WITHOUT MENTION OF HEPATIC COMA
ICD 9	070.54	CHRONIC HEPATITIS C WITHOUT MENTION OF HEPATIC COMA
ICD 9	070.7	UNSPECIFIED VIRAL HEPATITIS C
ICD 9	070.70	UNSPECIFIED VIRAL HEPATITIS C WITHOUT HEPATIC COMA
ICD 9	070.71	UNSPECIFIED VIRAL HEPATITIS C WITH HEPATIC COMA
ICD 9	V02.62	CARRIER OR SUSPECTED CARRIER OF HEPATITIS C
ICD 10	B17.1	ACUTE HEPATITIS C
ICD 10	B17.10	ACUTE HEPATITIS C WITHOUT HEPATIC COMA
ICD 10	B17.11	ACUTE HEPATITIS C WITH HEPATIC COMA
ICD 10	B18.2	CHRONIC VIRAL HEPATITIS C
ICD 10	B19.2	UNSPECIFIED VIRAL HEPATITIS C
ICD 10	B19.20	UNSPECIFIED VIRAL HEPATITIS C WITHOUT HEPATIC COMA
ICD 10	B19.21	UNSPECIFIED VIRAL HEPATITIS C WITH HEPATIC COMA
ICD 10	Z22.52	CARRIER OF VIRAL HEPATITIS C

Table S1 List of products used to define treatment for HCV.

Generic Product ID (10) DESCRIPTION	**MARKETED PRODUCT NAME**
BOCEPREVIR	VICTRELIS
DACLATASVIR DIHYDROCHLORIDE	DAKLINZA
ELBASVIR-GRAZOPREVIR	ZEPATIER
GLECAPREVIR-PIBRENTASVIR	MAVYRET
INTERFERON ALFA-2B	INTRON A
	INTRON A W/DILUENT
INTERFERON ALFACON-1	INFERGEN
LEDIPASVIR-SOFOSBUVIR	HARVONI
	LEDIPASVIR/SOFOSBUVIR
OMBITASVIR-PARITAPREVIR-RITONAVIR	TECHNIVIE
OMBITASVIR-PARITAPREVIR-RITONAVIR-DASABUVIR	VIEKIRA PAK
	VIEKIRA XR
PEGINTERFERON ALFA-2A	PEGASYS
	PEGASYS PROCLICK
PEGINTERFERON ALFA-2B	PEG-INTRON
	PEG-INTRON REDIPEN
	PEG-INTRON REDIPEN PAK 4
	PEGINTRON
RIBAVIRIN (HEPATITIS C)	COPEGUS
	MODERIBA

	MODERIBA 1200 DOSE PACK
	MODERIBA 800 DOSE PACK
	REBETOL
	RIBASPHERE
	RIBASPHERE RIBAPAK
	RIBATAB
	RIBAVIRIN
SIMEPREVIR SODIUM	OLYSIO
SOFOSBUVIR	SOVALDI
SOFOSBUVIR-VELPATASVIR	EPCLUSA
	SOFOSBUVIR/VELPATASVIR
SOFOSBUVIR-VELPATASVIR-VOXILAPREVIR	VOSEVI
TELAPREVIR	INCIVEK

91

92 3.2 List of predictors concepts

93 *Table S2 List of predictors used for creating features for the ML algorithm.*

PREDICTOR CONCEPTS
AGE
GENDER
RACE
ABDOMINAL CT SCAN
ABDOMINAL SURGERIES

ABNORMAL STOOL COLOR
ABNORMAL WGT LOSS
ACL INHIBITORS
ADDICTION MEDICINE SPECIALTY VISIT
ALCOHOL USE ABUSE DEPENDENCE
ALCOHOL WITHDRAWAL
ALCOHOLIC LIVER DISEASE
ALOPECIA AREATA
ALPHA 1 ANTITRYPSIN DEFICIENCY
AMBULATORY SPECIALTY VISIT
AMEBICIDES
AMINOGLYCOSIDES
ANALGESICS
ANOREXIA
ANTHELMINTICS
ANTI INFECTIVE AGENTS
ANTI INFLAMMATORY ANALGESICS
ANTI MOTILITY DRUGS
ANTI REJECTION AGENTS
ANTIANXIETY AGENTS
ANTIDEPRESSANTS
ANTIDIARRHEAL PROBIOTIC AGENTS

ANTIEMETICS
ANTIFUNGALS
ANTIHYPERLIPIDEMICS COMBOS
ANTIHYPERLIPIDEMICS MISC
ANTIMALARIALS
ANTIMYOBACTERIAL AGENTS
ANTIPHOSPHOLIPID SYNDROME
ANTIPSYCHOTICS ANTIMANIC AGENTS
ANTIRETROVIRALS
ANTIULCERANTS
ANTIULCERANTS PPIS
ANXIETY
ARTERITITS
ARTHROPOD BORNE HEMMORRHAGIC FEVER
ASCITES
AUTOIMMUNE HEMOLYTIC ANEMIA
B-CELL NON HODGKINS LYMPHOMA
BACTEREMIA
BARIATRIC SPECIALTY VISIT
BEHAVIORAL HEALTH SPECIALTY VISIT
BENIGN NEOPLASM
BILE ACID SEQUESTRANTS

BMT SCT TRANSPLANT
BRUISING
CACHEXIA
CARDIOPULMONARY BYPASS
CELIAC DISEASE
CEPHALOSPORINS
CHEST PAIN
CHLAMYDIA
CHOLANGITIS
CHOLESTEROL ABSORPTION INHIBITORS
CHOLESTEROL AGENTS
CHRONIC FATIGUE
CHRONIC LIVER DISEASE
CHRONIC LUNG DISEASE
CHURG STRAUSS SYNDROME
CIRRHOSIS
CKD ESRD
CLINICAL SOCIAL WORKER SPECIALTY VISIT
COLITIS
COLON CANCER SCREENING
COLONOSCOPY
COMPLETE BLOOD COUNT

CONFUSION
CONVULSIONS
COUNSELOR SPECIALTY VISIT
CRITICAL CARE SPECIALTY VISIT
CRYOGLOBULINEMIA
CYTOMEGALOVIRUS
DARK URINE
DEPRESSION
DIABETES
DIAGNOSTIC TESTING SPECIALTY VISIT
DIARRHEA
DIURETICS
DROWSINESS
DRUG SUBSTANCE WITHDRAWAL
DRY EYES
DYSARTHRIA
DYSMENORRHEA
DYSPEPSIA
DYSPNEA
EARLY SATIETY
EDEMA
EMERGENCY MEDICINE SPECIALTY VISIT

ENTERITIS DUE TO UNSPECIFIED VIRUS
EPIDEMIOLOGY PUBLIC HEALTH SPECIALTY VISIT
ERYTHROPOIESIS STIMULATING AGENTS ESAS
FAMILIAL HCV
FAMILY PRACTICE SPECIALTY VISIT
FEVER
FIBRATES
FIBROMYALGIA
FLU VACCINES
FLUOROQUINOLONES
GASTROENTEROLOGY SPECIALTY VISIT
GENERAL PRACTICE SPECIALTY VISIT
GENERAL SURGERY SPECIALTY VISIT
GENETICS SPECIALTY VISIT
GERD
GERIATRIC MEDICINE SPECIALTY VISIT
GLOMERULONEPHRITIS
GONORRHOEAE
GRANULOMATOSIS WITH POLYANGIITS
HCV TESTS
HEADACHE
HEART PALPITATIONS

HEARTBURN
HEMATOLOGY SPECIALTY VISIT
HEMATURIA PROTEINURIA URINALYSIS
HEMOCHROMATOSIS
HEMODIALYSIS
HEMODIALYSIS TREATMENT
HEMOPHILIA
HEMORRHOIDS
HEPATIC CARCINOMA
HEPATIC ENCEPHALOPATHY
HEPATIC FIBROSIS
HEPATIC OSTEODYSTROPHY
HEPATITIS CO INFECTION
HEPATITIS VACCINES
HEPATOLOGY SPECIALTY VISIT
HEPATOMEGALY SPLENOMEGALY
HERPES SIMPLEX VIRUS
HIGH RISK SEXUAL BEHAVIOR
HISTORY OF CARDIOPULMONARY BYPASS CABG
HIV AIDS
HOMELESSNESS ECONOMIC BURDEN
HUMAN PAPILLOMAVIRUS

HYPERGLYCEMIA
HYPERLIPIDEMIA
HYPERTENSION
HYPERTHYROIDISM
HYPOGLYCEMIA
HYPOTHYROIDISM
IBS
IMMUNE THROMBOCYTOPENIC PURPURA
IMMUNOLOGY SPECIALTY VISIT
IMMUNOSUPRESSIVES
INCARCERATION HISTORY
INFECTIOUS DISEASE SPECIALTY VISIT
INFECTIOUS MONONUCLEOSIS
INFLUENZA
INJECTABLE IRON
INSOMNIA
INSULIN RESISTANCE
INTERNAL MEDICINE SPECIALTY VISIT
IPF
JAUNDICE
JOINT PAIN
JUGULAR VEIN DISTENTION

LAB RESULT – ALT
LAB RESULT – AST
LAB RESULT – BILIRUBIN
LACTOSE INTOLERANCE
LICHEN PLANUS
LIVER ABSCESS
LIVER BIOPSY
LIVER DISEASE MULTIANALYTE ASSAYS
LIVER ELASTOGRAPHY
LIVER FAILURE
LIVER FUNCTION STUDIES
LOWER RESP TRACT INFECTION
LUPUS
LYMPHADENOPATHY
MACROLIDES
MAMMOGRAPHY
MICROSCOPIC POLYANGIITS
MILITARY SERVICE
MISC ANTI-INFECTIVE AGENTS
MOORENS CORNEAL ULCERS
MOSQUITO BORNE VIRAL ENCEPHALITIS
MTP INHIBITORS

MYALGIAS
NAUSEA VOMITING
NECROLYTIC ACRAL ERYTHEMA
NEUROPATHY
NICOTINIC ACID DERIVATIVES
NON-ALCOHOLIC STEATOHEPATITIS NASH
NON-HODGKIN LYMPHOMA
NON-INFECTIOUS HEPATITIS
NON0NARCOTIC ANALGESICS
NURSE PRACTITIONER SPECIALTY VISIT
OBESITY
OBSTETRICS GYNECOLOGY SPECIALTY VISIT
OCCUPATIONAL EXPOSURE
OPIOID ANALGESICS
ORGAN TRANSPLANT
OSTEOARTHRITIS
OTHER ABDOMINAL PAIN
OTHER ANTIVIRALS
OTHER DRUG USE ABUSE
OTHER FATIGUE
OTHER HEADACHE SYNDROMES
OTHER HEMORRHAGIC CONDITIONS

OTHER MALAISE
OTHER POXVIRUS INFECTIONS
OTHER PURPURA
OTHER STI
OTHER VASCULITIS
PCSK9 INHIBITORS
PENICILLINS
PEPTIC ULCER DISEASE
PHYSICIAN ASSISTANT SPECIALTY VISIT
POLYMYOSITIS DERMATOMYOSITIS
PORPHYRIA CUTANEA TARDA
PORTAL HYPERTENSION
PPIS
PROSTATE CANCER
PRURITUS
PSORIASIS
PSYCHIATRY SPECIALTY VISIT
PULMONOLOGY SPECIALTY VISIT
RAPE
RASH
RAYNAUDS PHENOMENON
REACTIVE ARTHRITIS

REGISTERED NURSE SPECIALTY VISIT
RENAL CANCER
RHEUMATOID ARTHRITIS
RHEUMATOID VASCULITIS
RIGHT SIDED HF
RIGHT UPPER ABDOMINAL PAIN
RISK OF INTRAVENOUS DRUG USE ABUSE
SCLERITIS
SCLERODERMA
SEDATIVES HYPNOTICS SLEEP DISORDER AGENTS
SENSORY NEUROPATHY
SJOGRENS DISEASE
SKIN ABCESS
SLOW VIRUS INFECTIONS
SLURRED SPEECH
SPIDER ANGIOMAS NEVUS
SPLENOMEGALY
STATINS
STD TESTS
STEATORRHEA
STEATOSIS
STREPTOCOCCUS PNEUMONIAE

SUBSTANCE USE ABUSE DEPENDENCE
SUBSTANCE USE DISORDER AGENTS
SULFONAMIDES
SWELLING OF LIMB
SYPHILIS
TETRACYCLINES
THALASSEMIA
THROMBOCYTOPENIA
THROMBOSIS
THYROIDITIS
TRANSEXUALISM
TRANSFUSIONS
TRANSVESTIC FETISHISM
TRICHOMONIASIS
ULCER THERAPY COMBOS
UNDERWEIGHT
UPPER ENDOSCOPY
UPPER RESPIRATORY TRACT INFECTION
URINARY RETENTION
UVEITIS
VACCINES (HEPATITIS or INFLUENZA)
VARICES

VIRAL CHLAMYDIAL INFECTIONS
VIRAL HEPATITIS
VIRAL PNEUMONIA
VITAMIN D DEFICIENCY
VITILIGO
WEAKNESS
XANTHELASMA XANTHOMA

94

95