Title

Exploring methods to identify individuals infected with hepatitis C virus in the United States: an application of ensemble learning with national survey data

Author

Telep, Laura E

Publication Date

2022

Peer reviewed|Thesis/dissertation

Exploring methods to identify individuals infected with hepatitis C virus in the United States: an application of ensemble learning with national survey data

By

Laura E. Telep

A dissertation submitted in partial satisfaction of the
requirements for the degree of

Doctor of Philosophy

in

Epidemiology

in the

Graduate Division

of the

University of California, Berkeley

Committee in Charge:

Professor Arthur L. Reingold, Chair
Professor Alan E. Hubbard
Professor Mahasin S. Mujahid
Professor Anand P. Chokkalingam

Spring 2022

Exploring methods to identify individuals infected with hepatitis C virus in the United States: an application of ensemble learning with national survey data

Abstract

Exploring methods to identify individuals infected with hepatitis C virus in the United States: An application of ensemble learning with national survey data

By

Laura E. Telep

Doctor of Philosophy in Epidemiology

University of California, Berkeley

Professor Arthur L. Reingold, Chair

Over 70 million people worldwide are living with chronic hepatitis C virus (HCV) infection. Untreated, HCV infection can progress to cirrhosis, advanced liver disease, and hepatocellular carcinoma. Our improved understanding of HCV transmission, coupled with significant advances in treatment have the potential to dramatically reduce the incidence and prevalence of HCV-related diseases. Based on these advances, the World Health Organization (WHO) has established a goal to eliminate HCV infection by 2030; however, a significant impediment to this goal is the lack of infection awareness, which perpetuates the spread of the virus. By improving the detection of HCV infection, we can connect patients to treatment to reduce its prevalence and curtail transmission to reduce future incidence of infection.

This dissertation reviews the literature on known risk factors for HCV infection in the United States (US) and uses a large, contemporary, publicly available national dataset, the National Health and Nutrition Examination Survey (NHANES), to look for additional risk factors and to build an algorithm to identify individuals with a high probability of HCV infection. NHANES participants are randomly selected from the non-institutionalized and housed US population and screened for HCV RNA, regardless of insurance status or known risk factors, providing meaningful insights into the characteristics associated with HCV infection.

The results of an umbrella review of circumstances associated with an increased prevalence of HCV infection in the US can be found in Chapter two. Risk factors were categorized as behavioral/lifestyle factors, risks associated with a medical condition, risks related to an occupation, or vulnerable populations. These findings can be used to improve outreach, education, and prevention programs, as many of the identified risk factors are present in marginalized groups that may not have access to regular healthcare or may be missed by existing HCV diagnosis and prevention efforts. Chapter three explores the use of ensemble learning methods to identify the features captured by NHANES that have the greatest impact on successful HCV infection prediction. NHANES data include HCV RNA measurements for all participants of the medical examination portion of the survey. With this information, the ensemble learning method Super Learner was used to identify complex patterns of characteristics

associated with HCV infection and to identify the characteristics that had the greatest impact on successful HCV infection prediction in the US (ranked variable importance). Using a subset of the NHANES data that would likely be available and accurate in electronic medical records, Chapter 4 examines the development of an HCV prediction algorithm that could be used to prioritize candidates for HCV screening. Overall, these findings contribute to the national effort to increase HCV-infection detection and accelerate progress towards HCV elimination.

Table of Contents

Dedication

In memory of Maria Alvarado. We all miss you.

Acknowledgements

I was incredibly fortunate to participate in an internship at Gilead Sciences during my MPH training at Berkeley in 2014. Through this internship I had the privilege to work with Dr. Anand Chokkalingam, who subsequently agreed to teach me through independent study when my schedule become unmanageable; helped me identify a project and became my masters project advisor; then became my supervisor and research collaborator at Gilead; and now serves as one of my advisors in the doctoral program. Anand has offered me amazing opportunities, mentorship, support, and genuine kindness, and I am so very grateful for all of this.

I am also grateful to have had the support of the absolutely brilliant and remarkably kind faculty in the School of Public Health at Berkeley, who enthusiastically took on the challenge of training me, despite the fact that we were in the middle of the pandemic when they were juggling their own challenging personal circumstances. Thank you to my advisor, Art Reingold, for his invaluable wisdom, his boundless generosity, and his lifetime commitment to public health. I will always be indebted to Patrick Bradshaw whose enthusiasm for epidemiologic methods is completely infectious. Profound thanks to Rachael Phillips, Andrew Mertens, Jeremy Coyle, and Ivana Malenica, who guided me through the analytic aspects of my dissertation with patience and cheerfulness. Thank you to Mika Pejovic, who took a break from teaching IB Biology to explain the intricacies of viral replication to me and to Jennifer Head, Shelley Facente, Shalika Gupta, Emon Elboudwarej, and Moon Choi, whose mock qual prep gave me confidence that I might survive the QE.

I was so lucky to have amazing committees through my doctoral journey. In particular, I am indebted to Drs. Art Reingold, Anand Chokkalingam, and Alan Hubbard for agreeing to serve on my dissertation committee, to Drs. Sandi McCoy, Maya Peterson, and Eva Harris for serving on my advancement to candidacy exam, and to Dr. Mahasin Mujahid for doing double duty by serving on both.

I must express my gratitude for the amazing cohort of epidemiologists who started in the epidemiology doctoral program with me in 2019. They are intelligent, passionate and compassionate, and I fully expect they will improve health outcomes wherever they choose to direct their energy. I am humbled and slightly shocked that I was allowed to train with them (their organization for QE prep was absolutely mind-boggling) and I am unquestionably a better epidemiologist because of them.

Thank you to my husband Chip, who has been unendingly patient and who brags about me to anyone who will listen, and my children Margaret and Claire – who were in middle school when I started in the MPH program and are now juniors in college. They have endured my endless Zoom calls and weekends on the computer, and believed in me every step of the way, even when I didn't. They made me laugh and helped me keep the drama in perspective. They have taught me so much about how to be supportive and about unconditional love. I can't believe how lucky I am to call them my family.

List of Figures

List of Tables

Chapter 1: Introduction

Over 70 million people worldwide are living with chronic hepatitis C virus (HCV) infection[1]. HCV is a genetically diverse, parenterally transmitted virus that causes inflammatory liver disease and, untreated, can progress to liver decompensation and hepatocellular carcinoma (HCC) [2, 3]. In 2017, 3,621 acute HCV cases were reported to the Centers for Disease Control and Prevention (CDC), reflecting a rate of 1.2 cases per 100,000 population nationally - with the highest incidence rates seen in Indiana and West Virginia (4.0 and 3.9 cases per 100,000, respectively). After adjusting for under-reporting and under-ascertainment, it is estimated that there were 50,300 (95% confidence interval (CI): 39,800 – 171,600) new (acute) HCV infections in the United States that year, and approximately 2.4 million people living with chronic HCV infection [4]. Percutaneous exposure to contaminated blood is the primary mode of transmission of HCV, most commonly due to injection drug use and unsafe medical practices, and less frequently through sexual contact and perinatal transmission [5, 6]. In 2012, the prevalence of HCV infection in the United States among those born from 1945 to 1965 was as much as five times higher than that in other age groups[7, 8]. Today the incidence of HCV infection is growing fastest in younger adults, driven by the opioid epidemic [9], with 3.1 and 2.6 acute cases of HCV infection per 100,000 in 20 and 30 year-olds respectively, in 2018, compared to 0.9 and 0.4 new cases per 100,000 among people in their 50s and 60s [4].

There is no vaccine to prevent HCV infection. After initial infection, it is estimated that 30% (15-45%) of those infected will spontaneously clear the virus within the first six months [10]. Those who develop chronic HCV infection have a 15-30% risk of developing cirrhosis within 20 years [11, 12], and those with HCV and cirrhosis have a 3% annual risk of HCC, representing a 15-20-fold increased risk over those not infected with HCV. Interventions prior to the onset of cirrhosis represent the greatest opportunity to change the clinical course of disease [13].

In the past, treatment of HCV relied on interferon-based regimens, which required almost a full year of weekly injections; were associated with important side-effects; and had limited success in achieving a sustained virologic response [14]. Safe and effective direct acting antivirals (DAAs) that can be administered orally became available for treating HCV in 2014. With their minimal side effects and comparatively short course of treatment, DAAs dramatically changed the HCV treatment landscape, and created a real opportunity to reverse the growing incidence and prevalence of HCV infection and related disease[15, 16]. Inspired by the success and availability of both the hepatitis B virus (HBV) vaccine and of HCV treatment, the World Health Organization (WHO) has challenged health agencies around the world to take action against viral hepatitis by establishing a sustainable development goal to eliminate hepatitis B and C. Specific targets include a 90% reduction in HCV incidence and a 65% reduction in HCV-related mortality by 2030 [17].

Eliminating HCV will require interventions on many levels, including harm reduction measures, such as needle exchange programs and education; increased surveillance and screening; and linking those who test positive to treatment. Infection awareness presents one of the biggest obstacles in the HCV cascade of care. Published findings indicate that up to 80% of persons with HCV infection are unaware of their infection, including over 50% in the United States [4, 18, 19]. In addition, it is estimated that only one out of every 12 HCV infections in the United States

is reported to the National Notifiable Disease Surveillance System (NNDSS) [20]. Impediments to accurate reporting include cases not being recognized in the absence of symptoms; patients not seeking medical attention, even when symptomatic; and physicians not reporting due to delayed test results or problematic case definitions [20, 21]. Poor access to screening and healthcare in some high-risk populations also contributes to under-reporting [22]. However, even among those with sufficient healthcare access, it is believed that the majority of HCV infections go unrecognized and undiagnosed [23]. This undiagnosed population represents a missed opportunity to both reduce the prevalence of HCV and reduce future incidence by minimizing opportunities for transmission.

HCV Screening Recommendations

The goal of screening is to identify individuals likely to have a disease or infection as early as possible - prior to symptom onset - in hopes of altering the disease course and improving outcomes. HCV screening in the United States has evolved along with the prevalence of HCV infection and demographic characteristics of those with the infection. In 2004, the USPSTF explicitly recommended against screening for HCV in asymptomatic adults, a grade D recommendation indicating that screening had no net benefit. While screening was found to be effective at identifying HCV infection, the prevalence of HCV infection was considered low in the general population, and there was no evidence that early detection would improve long term health outcomes [24].

In 2013, HCV infection and related chronic liver disease had become the most common indication for liver transplantation[25] in the United States, and the primary driver of HCC[2, 26], prompting the USPSTF to upgrade its HCV screening recommendations to a grade B, which acknowledged moderate to substantial benefits. Importantly, B grades also ensure that insurance will cover the cost of screening with no deductible or copay under the Affordable Care Act of 2010 [27]. The updated recommendations targeted individuals 18-79 years of age at high risk for HCV infection, in addition to recommending one-time screening for all adults born in the birth cohort from 1945 and 1965 in whom the prevalence of HCV was between 2.6-3.5%[8, 28]. Factors associated with a high risk of HCV infection included a history of drug use, incarceration, homelessness, high risk sexual behavior, blood transfusions that occurred prior to 1992 (when screening of blood for HCV was implemented), hemodialysis, laboratory results indicating elevated liver enzymes, and maternal HCV infection [29]. Studies showed that after these recommendations were put into place, screening in the 1945 – 1965 birth cohort increased significantly, but little change occurred in other age groups[30].

Evidence suggests that broader screening recommendations than those currently in place are warranted, given the inadequacy of exclusively risk-based approaches to contain the increasing incidence of HCV and the positive impact that such changes can have on screening practices [30, 31]. A 2015 study by the European Union (EU) HCV Collaborators modeled the impact of existing interventions on the prevalence of HCV infection at time points over a 15-year period to determine whether existing measures would be sufficient for the EU to achieve the WHO 2030 elimination targets. They estimated 3.2 million people had active HCV viremia in Europe in 2015, of whom, 36% had been diagnosed, 5% had been treated, and only 4% had been cured. In the same year, they estimated an additional 58,000 people had been newly infected with HCV,

and over 30,000 people with HCV infection immigrated to the EU. By maintaining the 2015 screening and treatment scenarios, the model projected that the entire diagnosed population would be cured by 2030, while the population with HCV viremia would decline by only 40%, falling well short of the WHO elimination goals. To achieve a 90% reduction in incidence of HCV infection, they estimated that annual HCV diagnoses would need to increase two-fold, and screening would need to expand significantly to achieve these targets [32]. These findings highlighted the need to expand screening from only high-risk patients to those in the general population.

In the United States, the CDC and the US Department of Health and Human Services (HHS) sponsored an Institute of Medicine committee (the Committee on a National Strategy for the Elimination of Hepatitis B and C) to explore the feasibility of hepatitis elimination and devise a path forward. The two-volume report of their findings was published in 2017. Among the obstacles to elimination they identified are poor awareness of HCV infection and limited resources devoted to education and outreach, noting that, at the time of publication of the report, CDC funded only seven jurisdictions for viral hepatitis surveillance. [31]. The report indicated that 250,000 people with HCV infection will need to be diagnosed and treated annually for the United States to reach elimination goals and suggested that one-time universal screening of all adults would be a necessary step to identify a significant portion of the infected-but-unaware population. This estimate is supported by a modeling study conducted by Kabiri et. al (2014) that projected HCV infection rates from 2001 – 2050 using NHANES data under different screening scenarios. Study findings suggested that over the next 10 years, universal HCV screening would identify twice as many HCV infections as risk-based screening – up to 450,000 additional HCV-infected persons [33].

In March 2020, driven by the WHO elimination goals, the significant rise in reported HCV infections related to injection drug use and the opioid crisis, the effectiveness of DAA treatment regimens, and the overwhelming evidence that existing guidelines were failing to capture a significant proportion of the HCV-infected population, USPSTF updated its screening recommendations. The new guidance includes one-time screening for all adults 18-79 years of age, including pregnant women, and more frequent screening for those with identified risk factors, including adolescents [34]. These recommendations have maintained their grade B status.

Screening Challenges

Although screening recommendations have now been expanded, adoption of these new guidelines has been slow [35-37]. There remain major financial, logistical, and social obstacles to uptake and implementation. A review examining barriers to HCV screening looked at studies published between 2012 and 2017 and identified limiting factors on both the patient and the provider side. For patients, low self-perceived risk, fear of testing positive, and perceived stigma associated with HCV were cited as reasons to decline screening. For providers, barriers included lack of HCV-specific knowledge, time constraints in the face of competing patient needs, and discomfort inquiring about potential HCV risk factors[38]. A recent study of HCV-infected baby boomers who had received treatment at the Penn Hepatology clinic and who had adult children found that more than half were unaware that HCV could be transmitted vertically, and many were unwilling to consider that they may have been infected prior to delivery. Participants were

offered the option to have their children receive HCV education and free testing and treatment through the study. Fewer than half of study participants agreed to let their children be contacted, and among those children who were contacted, only half agreed to be tested for HCV[39].

Alter et al. pointed out that expanding HCV testing and treatment will require a significant financial investment up front, even in comparatively wealthy countries such as the United States [40]. Studies examining the economic impact of increased HCV testing and treatment have demonstrated that the costs to screen and treat patients in the short term will be substantial, although ultimately far less than the long-term costs associated with the cirrhosis, liver cancer, and liver transplantation that would potentially occur without treatment [41-44]. A study by Scott et al. modeled the direct and indirect economic impacts of scaling up HCV testing and treatment to achieve the WHO 2030 elimination goals. Their findings suggest that globally, an initial investment of almost $5 billion would be needed, but this investment would prevent 2.1 million HCV-related deaths and 10 million new HCV infections, producing a net economic benefit of $22.7 billion by 2030 [45].

Priorities in scaling up HCV screening

Experience from the campaigns to eliminate smallpox and polio points to the critical roles of surveillance and screening to control and ultimately eliminate a disease [46-48], and the importance of using data strategically to monitor progress and target efforts. Given competing needs and the finite capacity of our health systems, we need to be strategic about how resources can be most effectively deployed to maximize identification of people infected with HCV. A report published in 2019 from the Coalition for Global Hepatitis Elimination highlighted the need to use strategic data to prioritize activities and guide the distribution of resources, noting that these are essential components of disease elimination programs [49]. Recent studies have described strategies to expand the reach of screening programs while identifying those patients who would receive the greatest benefit from being prioritized for testing. Approaches that have shown success in screening for HCV include electronic reminders in patient records to identify screening candidates [50], professional education and support for non-specialists to identify and manage HCV infection [51], and the development of tools to leverage electronic health records to identify candidates at increased risk of HCV infection [51, 52]. For example, a committee of HCV specialists led by Dieterich at Mount Sinai Hospital developed an algorithm to assist non-specialist health care providers in identifying and managing patients with uncomplicated HCV infection, and in recognizing those individuals who should be referred to specialists [53]. Tools such as these can be strategically employed to maximize the opportunities to identify individuals infected with HCV.

There is an opportunity in this moment to capitalize on the momentum of expanded HCV screening recommendations and the WHO goal of eliminating HCV as a public health threat by 2030. Prioritizing for screening those individuals most likely to be HCV-infected will yield the biggest impact on the prevalence of infection and the subsequent transmission of HCV infection and is the best use of finite resources as we scale up screening efforts. To this end, this dissertation summarizes the known risk factors for HCV infection (Chapter two) and examines strategies to build on this knowledge (Chapters three and four). The ensemble learning method Super Learner is used to identify potentially novel HCV-infection risk factors and to build an algorithm that could be used to prioritize high-risk candidates for HCV screening.

Chapter 2: Umbrella review of risk factors associated with hepatitis C virus infection in the United States

2.1: Abstract

Hepatitis C virus (HCV) infection represents a significant public health problem worldwide. As a result, eliminating hepatitis as a major public health threat is one of the goals proposed in the 2030 Agenda for Sustainable Development by the World Health Assembly. Targets of this goal include diagnosing 90% of individuals infected with HCV and treating 80% of the treatment eligible population. In the United States (US), a significant impediment to achieving these targets lies in the substantial under-diagnosis of HCV infection, due in part to the asymptomatic nature of early-stage infection. To bridge this gap in diagnosis, it is critical to understand the risk factors associated with increased prevalence of HCV infection. In this umbrella review, systematic reviews or meta-analyses of observational studies reporting HCV antibody prevalence were identified from inception to December 2021 using BIOSIS, Embase, MEDLINE®, and Cochrane bibliographic databases, and manual reference screening.

In the 13 reviews that met the inclusion criteria, eight risk groups were identified in the US in which HCV infection prevalence was disproportionately higher than that in the general population. They included three behavioral/lifestyle factors, two medical conditions, one occupation, and two vulnerable populations. HCV infection prevalence estimates among groups with multiple risk factors were substantially higher than HCV prevalence in the general population (~1.4%) or in individuals with only one identified risk factor. The highest pooled prevalence estimates occurred among men who have sex with men (MSM) who have ever injected drugs (29.9%; 95% confidence interval (CI): 16.5 – 45.2%), HIV-positive MSM who inject drugs (35.6%; 95% CI: 21.1 – 50.1%), and adult incarcerated individuals in North America with a history of injection drug use (67%; 95% CI: 25 – 80%). Across all included reviews prevalence estimates from individual studies that contributed to the systematic reviews ranged from 0% in healthcare workers to 84.9% in people who inject drugs, illustrating the significant heterogeneity in included studies.

Identifying risk factors associated with HCV infection may assist in developing targeted outreach and prevention strategies for populations at high risk of exposure to HCV and serve as a starting point to identify additional risk factors.

2.2: Introduction

Hepatitis C virus (HCV) infection occurs through parenteral exposure to infected blood[54]. The virus preferentially infects hepatocytes and is associated with significant morbidity and mortality, making it a major public health problem around the world. The availability of highly effective direct acting antiviral drugs (DAAs) has changed the landscape regarding HCV infection and resultant disease, making the elimination of HCV an achievable ambition[15, 16, 55]. To this end, the WHO has challenged global health agencies to prioritize viral hepatitis (B and C) elimination by 2030[56]. Achieving this goal, however, presents a significant challenge, as it will require the identification and treatment of at least 80% of people infected with hepatitis B virus and HCV to address current infection and prevent future transmission.

Based on data collected from 2013 – 2016 for the National Health and Nutrition Examination Survey (NHANES) and sources used to capture individuals not included in NHANES (e.g. nursing home residents, military personnel, incarcerated persons and unhoused individuals), an estimated 4.1 (3.4 – 4.9) million adults (1.7%) in the United States (US) were HCV antibody positive and 2.4 million were HCV RNA positive[57]. In addition, over 18,000 deaths reported to the National Vital Statistics System in 2016 were attributed to HCV-related sequelae. These estimates are considered conservative due to under-reporting and under-diagnosis of HCV infection[58]. According to the Centers for Disease Control and Prevention (CDC), the number of reported acute HCV infections increased every year from 2009 to 2019, with the highest incidence rates among persons 20-39 years of age[59]. Limiting the spread of HCV is particularly challenging in that the initial stages of infection are often asymptomatic, allowing the virus to persist and be passed on to others undetected. In the cascade of HCV care, identifying the undiagnosed, HCV-infected population remains the largest challenge[60]. It is estimated that up to 50% percent of individuals infected with HCV in the US are unaware of their infection[60, 61].

To bridge this gap in diagnosis and support elimination efforts, the United States Preventative Services Task Force (USPSTF) has expanded HCV screening recommendations to include universal one-time testing for HCV antibodies for all adults (>= 18 years of age) in any setting where the prevalence of HCV infection is >= 0.1%[62]. This important step is one of many proactive efforts needed to identify the undiagnosed HCV-infected population. Universal screening is necessary, but likely not sufficient to meet the WHO HCV elimination targets. Additional innovative efforts will be necessary. While government agencies around the world have taken a variety of approaches to identifying and treating those infected with HCV, there are financial and logistical challenges and competing priorities that hinder this effort[63, 64].

In support of efforts to locate the individuals who are, both knowingly and unknowingly, infected with HCV, it will be important to understand the conditions associated with the highest prevalence of HCV infection. To this end, we have conducted an exhaustive literature review to summarize the known risk factors and populations associated with an increased prevalence of HCV infection in the US. This review focuses exclusively on systematic literature reviews (SLRs) to clarify existing knowledge of HCV risk factors and to identify gaps that should serve as a starting point for future research into novel HCV risk factors and high-risk populations. To our knowledge, no umbrella review has previously been conducted that summarizes HCV infection prevalence in US SLRs.

2.3: Methods

Search Strategy

Systematic literature reviews containing estimates of the prevalence of HCV and the odds or risk of HCV infection in general and in high-risk populations were retrieved from BIOSIS, Embase, MEDLINE®, and Cochrane using Ovid®. The initial search strategy used broad terms to describe HCV infection and SLRs (a full list of search terms employed, and bibliographic databases searched is provided in **Supplemental Tables S2.1** and **S2.2**). The search was conducted on December 5, 2021, with no restrictions placed on date or geographic region. SLRs that were presented only in conferences and meetings were excluded. A manual search of the

references of the identified SLRs was performed to identify additional relevant publications for inclusion.

Study selection and eligibility
After the initial search using the Ovid® medical research platform, titles and abstracts of identified studies were screened for relevance. A full text screening was then performed on the relevant studies. Two types of SLRs were retained – those that focused on a particular HCV infection risk factor and those that focused on an epidemiologic feature of HCV infection in a specific geographic region and included analyses of high-risk subpopulations. Studies were excluded for the following reasons:

- not an SLR
- not written in English
- not conducted on a US population
- HCV infection was the risk factor and not the outcome
- did not include discussion of observable HCV risk factors
- didn't clearly describe methods or inclusion/exclusion criteria
- incomplete or unlisted references
- did not include prevalence information for HCV risk factors

The retained SLRs were categorized by identified HCV risk factors, and the references of individual SLRs were explored. If multiple reviews contained the same individual study references, the review with the most comprehensive and contemporary reference list was retained.

Data extraction
Because the goal of this umbrella review was to cast as wide a net as possible to identify any possible HCV infection risk factors, we focused on HCV antibody prevalence (not on HCV RNA). For each included SLR, the following information was extracted, if available:

- first author's name
- year of publication
- data cutoff of included literature
- risk factors described
- included study designs
- number of studies
- total number of participants
- prevalence measures and heterogeneity assessment.

For studies that characterized multiple risk factors for HCV infection, an assessment of each risk factor was conducted separately. Only risk factors reported with sufficient detail (i.e. itemized references of studies with at least 50 people) were included in this review. The total numbers of studies and participants for each risk factor were identified and listed separately.

Data synthesis
Given the heterogeneity of outcome measures, reporting styles, and HCV risk factor definitions, and the inherent difficulty in combining findings from systematic reviews of observational data, a meta-analysis was not performed. Instead, a narrative synthesis for each HCV infection risk

factor in the US was conducted, including all relevant and available measures of effect size (pooled measures, or ranges when a pooled measure wasn't available). HCV risk factors were organized and synthesized based on the following categories: behavioral and lifestyle risk factors, medical conditions or procedures, occupations, and special populations.

2.4: Results

A total of 3,099 studies were identified in the primary search of the referenced bibliographic databases: Medline: 990, BIOSIS: 952, Embase: 1022, Cochrane: 135 (see appendix for a full list). Conference and meeting reviews and duplicate references were excluded, leaving 1,161citations. Screening of titles and abstracts excluded 1,006 additional reviews. Among the remaining 155 SLRs, a manual investigation of bibliographies was conducted in which two additional reviews were identified. Finally, a full text review was conducted by individual risk factor to ensure the uniqueness of references and quality of content. In addition, only references and reviews of US data were retained. This last step excluded 144 SLRs, yielding a final total of 13 included SLRs (see **Figure 2.1**). A summary of the characteristics of the included reviews organized by risk factor is provided in **Table 2.1.** Any SLR that examined more than one HCV risk factor was listed with each risk factor grouping, so a review may appear more than once in **Table 2.1**. A summary table organized by review (n = 13 rows) that itemizes all risk factors and total numbers for included studies and participants is available in **Supplemental Table S2.3**.

HCV infection risk factors identified in this review have been grouped into four categories: (1) behavioral/lifestyle, (2) medical conditions/procedures, (3) occupations, and (4) special populations. The behavioral/lifestyle category includes people who inject drugs (PWID), non-injection drug users (non-IDU), and men who have sex with men (MSM). The medical conditions/procedures category includes human immunodeficiency virus (HIV) infection and severe mental illness. The occupations category includes health care workers (HCW). Last, the special populations category includes persons who are incarcerated, or homeless. Below is a summary of US-specific findings for each risk factor in each category.

Behavioral/Lifestyle Risk Factors

<u>People who inject drugs (PWID)</u>

Eight SLRs published from 2013 to 2021 included 35 US studies with a total of more than 7,000 participants investigating the prevalence of HCV infection among PWID[65-72]. Two studies reported pooled prevalence measures ranging from 29.9% (16.5 – 45.2%) among MSM who have ever injected drugs[72] to 35.6% (21.1 – 50.1%) among HIV-positive MSM who inject drugs[66].

Six reviews did not include pooled measures but in individual studies, HCV infection prevalence measures ranged from 17.9 to 84.9%, with the highest prevalence occurring in studies of PWID who have other HCV infection risk factors, such as homelessness and incarceration[65, 67, 71]. Wirtz et al. reviewed HCV prevalence among incarcerated persons in the US and included six studies that reported an HCV infection prevalence among incarcerated PWID ranging from 35.8 to 84.9%. A review of risk factors for HCV among drug users by Zhou et al.[69] included two US-based studies. Among individuals in the general population in Tennessee offered opt-out

HCV testing at family planning and sexually transmitted infection (STI) clinics, 48% (271/576) of self-reported PWID tested positive for HCV antibodies[73], while in a second study among self-reported injection drug users in Baltimore, 82.2% (204/248) tested positive for HCV antibodies [74].

Two SLRs examined HCV infection prevalence in US populations that have been disproportionately affected by HCV. Bruce et al. examined HCV infection prevalence among American Indians and Alaska natives (AI/AN). In four studies that reported HCV infection prevalence among AI/AN PWID, the prevalence ranged from 25.7 to 67.6%[68]. Schalkoff et al. examined drug use and associated infections in rural Appalachia. They included three studies that report HCV infection prevalence among PWID, with measures ranging from 34.0 – 54.8% and cite injection of cocaine and sharing of injection equipment to be correlated with the increased prevalence[70].

Non-injection drug use (non-IDU)

One SLR by Jordan et al. included US data on the prevalence of HCV infection among HIV-positive MSM who use non-injection drugs. From six studies with a total of 3,187 individuals, they reported a pooled HCV infection prevalence of 7.5% (5.2 – 9.9%). The prevalence reported in individual studies ranged from 4.9 to 12.3% [66], with a higher infection prevalence of HCV infection reported in the more recent studies.

Men who have sex with men (MSM)

We identified two SLRs with US data that examined the prevalence of HCV infection in MSM. A review by Jin et al. included 21 US studies with a total of 29,523 participants and reported a pooled HCV infection prevalence of 4.4% (2.6 – 6.1%) among MSM. When stratified by HIV infection status, the pooled HCV prevalence in HIV-positive MSM and in HIV-negative MSM was 9.7 % (7.1 – 12.6%, n=12 studies) and 2.3% (0.7 – 4.8%, n = 7 studies), respectively[72]. A second review by Wirtz et al. included one study conducted at a correctional facility in California that reported a 50% prevalence of HCV infection among incarcerated MSM[67].

Medical Conditions/Procedures

HIV Infection

Two global SLRs included US data on the prevalence of HCV infection among HIV-positive individuals. A review of HCV infection in HIV-positive MSM by Jordan et al. included seven US studies with a total of 4,074 individuals. The overall pooled prevalence of HCV infection in these studies was 12.0% (8.5 – 15.4%) among HIV-positive MSM, with individual study prevalence measures ranging from 6.0% to 17.9%[66]. A second global review by Jin et al. included 12 US studies with a total of 11,502 HIV-positive MSM and reported a pooled HCV infection prevalence of 9.7% (7.1 – 12.6%) in this group[72].

Severe Mental Illness

HCV infection prevalence in the US among people with severe mental illness was characterized in two SLRs. A 2018 review by Ayano et al. included two studies that reported the prevalence of HCV infection to be 5.9% (252/4310) and 16.1% (122/777), respectively in individuals with severe mental illness. When stratified by sex, the HCV infection prevalence in men with severe

mental illness was more than double that in women in both studies (6.2% for men vs. 2.3% for women in the first study and 19.8% for men vs. 9.8% for women, in the second)[75], a ratio consistent with the higher prevalence of HCV infection in men in the general population[76]. The second SLR by Hughes et al. published in 2016 included 12 US studies with a total of 4,977 individuals. They estimated the pooled prevalence of HCV infection among people with severe mental illness in the US to be 17.4% (95% CI: 13.2 – 22.6), with individual study prevalence measures in this population ranging from 2.7 to 38%[77].

Occupations:

Healthcare Workers

Data on HCV infection among healthcare workers (HCW) in the US were available in one SLR that included studies from around the world. This review included eight US studies of HCW published between 1991 and 2007 and including a total of 7,213 individuals; the HCV infection prevalence estimates ranged from 0% to 1.9%, consistent with the prevalence in the US general population[78]. Of note, they also reported a pooled odds ratio of 2.1 (95% CI: 1.3 – 3.42, I^2=70) when comparing HCV infection among healthcare workers in the US and Europe to controls (blood donors) from those countries. The prevalence of HCV infection was found to be higher in older studies than more recent ones[79].

Special Populations

People who are incarcerated

Three SLR included US data characterizing HCV infection prevalence among incarcerated individuals. The reviews were published between 2013 and 2018 and contained 35 individual studies of over 2.2 million individuals combined. All reviews provided pooled prevalence estimates of HCV infection among incarcerated persons in North America that were well above the US and Canadian general population HCV prevalence estimate of 1.5%[78].

Larney et al. estimated a pooled HCV infection prevalence of 29% (95% CI: 24 - 35%, I^2 = 98.9) for adult incarcerated persons in North America, combining nine studies from the US (n = 19,223) and six from Canada (n = 6,189). The US studies reported HCV infection prevalence measures that ranged from 20.7 to 50.5%[65]. Dolan et al. estimated a pooled HCV infection prevalence of 15.3% (95% CI: 13.1 - 17.7%, I^2 = 99.9) in North America, with 20 US studies that included 35 infection prevalence estimates ranging from 0.8% to 41.4%, and two Canadian studies with three estimates ranging from 17.6% to 25.2%[80]. The third review by Wirtz et al. estimated a pooled HCV infection prevalence of 11.3% (95% CI: 6.3 - 20.3%, I^2 = 96.4) in North America that included five US estimates ranging from 0.1% to 17% in incarcerated non-injection drug users, and two Canadian estimates ranging from 2.7 to 4.1%[67]. In both the Dolan et al. and Wirtz et al. reviews, the lower US prevalence estimates were in juvenile detention centers.

Homeless Individuals

One global review of HCV infection prevalence in homeless individuals by Beijer et al. (2012) included data from five US studies (combined n = 1,758). HCV infection prevalence estimates stratified by sex ranged from 8% in males and 10% in females in a 2009 study conducted at a

homeless shelter in Honolulu, to 35% in males and 27% in females in a 2001 study of homeless individuals who engaged with a mobile clinic in New York City[81].

2.5: Discussion

The goal of this umbrella review was to synthesize evidence concerning known risk factors for HCV infection in the US. This review can serve as a starting point for efforts to diagnose and treat individuals infected with HCV in support of the WHO's 2030 global hepatitis elimination goals. Evidence in the review comes from two types of SLRs: those that characterized the prevalence of HCV infection among individuals with specific risk factors and those that described the epidemiologic features of HCV infection in a specific US region or population subgroup and included prevalence data on risk factors for infection in that group.

In the 13 SLRs that met our inclusion criteria, we identified eight risk groups in the US in which HCV infection prevalence was disproportionately higher than that in the general population. They included three behavioral/lifestyle factors, two medical conditions, one occupation, and two vulnerable populations. Because few of the identified reviews assessed HCV RNA (current HCV infection), this review focused on HCV antibody prevalence, which captures individuals with either current or past HCV infection.

It is important to note that many of the identified risk factors summarized in this review are concentrated in vulnerable and marginalized groups in the US. Given the availability of well-tolerated and effective treatment options, we have the opportunity to eliminate HCV infection; however, elimination will require health agencies to target resources to groups that are often underserved - people who are homeless, incarcerated, suffer from mental illness, or inject drugs. Without a genuine investment in these communities, it is unlikely that elimination will be achieved.

Several SLRs of HCV infection risk factors and high-risk populations were identified for PWID (eight reviews) and for incarcerated persons (three reviews). Two SLRs were identified for each of the following groups: people with severe mental illness, people infected with HIV, and MSM. We identified one review of US data for each of the remaining risk groups: people who use non-injection drugs, homeless individuals, and healthcare workers (see **Table 2.2**). It is important to remember that many additional SLRs were identified in our initial search of the bibliographic databases. When more than one review included the same studies, we kept only the review that had the most comprehensive and contemporary references.

The eight systematic reviews of PWID consistently reported significantly higher prevalences of HCV infection in this population. This evidence provides insights into the increase in HCV incidence in young people[59]. After the initial identification of HCV in 1989[82, 83], the prevalence of the infection was concentrated in people born between 1945 and 1965 ("baby boomers"), due to previous high-risk behaviors and encounters with the medical establishment before preventative measures were established, such as screening of donated blood[84]. In 2013, the USPSTF singled out baby boomers for additional surveillance, including a one-time screening for HCV antibodies[8, 29]. The rise of injection drug use has markedly changed the age distribution of the HCV-infected population. A systematic review done in 2020 by Hines et al. reported the mean age of PWID in the US to be 37.7 years of age (28 – 55), with more than

10% of PWID aged 25 years or younger[85]. The impact of opioid addiction has led to the bimodal concentration of HCV infection in individuals aged 50-69 years and in those aged 20-39 years[86].

Non-injection drug use, also an HCV risk factor, has a less clear mechanism for transmission of HCV infection. Only one review met our inclusion criteria, and its focus on HIV+ MSM suggests alternative pathways for HCV infection. An SLR by Scheinmann et al. identified methodologic concerns (e.g. misclassification of drug use, lack of a primary focus on this population, etc.) that may limit our understanding of the association between non-IDU and HCV infection, and suggests that because the prevalence of HCV infection in this group continues to be higher than that in non-users of drugs, further research is warranted [87].

Three systematic reviews addressed the increased prevalence of HCV infection in incarcerated persons, and one systematic review included studies that characterized the prevalence of HCV infection in homeless individuals. A review by Degenhardt et al. reported that both homelessness and incarceration are also strongly association with injection drug use [88]. It is notable that several SLRs of HCV infection among PWID examined the risk within these two communities, illustrating that HCV infection risk is likely to be multifactorial. It is possible that injection drug use is instrumental in several other identified risk factors as well (e.g. severe mental illness, HIV infection, and MSM). A review by Schiffman et al. posited that preventing infection and re-infection with HCV in the PWID community would require interventions that go beyond HCV treatment to address underlying medical and social factors that put them at greater risk[89].

Estimates of HCV infection prevalence in US MSM and in HIV-positive individuals were often combined. The review by Jin et al.[72] looked at HCV infection prevalence in MSM overall and stratified by HIV infection status. While the prevalence of HCV infection was higher in the HIV-positive group than in the HIV-negative group, the seven studies with a total of 7,362 HIV-negative MSM still had a pooled prevalence of HCV infection higher than that in the general population, providing evidence that MSM may have an elevated risk of HCV infection. The review by Jordan et al.[66] examined HCV infection in HIV-positive MSM who were drug users. After stratifying by injection drug use, HCV infection prevalence estimates were higher in HIV-positive MSM who inject drugs, and lower in HIV-positive MSM who use non-injection drugs, but all included studies reported prevalence estimates higher than that in the general population.

Progress has been made in limiting HCV infection risk in some circumstances. While healthcare workers in the US and other middle and high income countries previously experienced a small increased risk of HCV infection through occupational accidents (contaminated sharps, etc.), that risk has declined over time through the development of robust infection prevention protocols[90] and improved blood screening[91].

This review has several strengths. This is the first umbrella review to bring together SLRs of known risk factors for HCV infection in the US. We have summarized the findings from 13 reviews that include 105 US studies with over 2.3 million individuals to thoroughly describe the known landscape of HCV infection risk in the US. The review findings are important because they illustrate the multifactorial nature of HCV infection and provide insights into pathways where interventions could have the biggest impact.

In the US, many of the high-risk groups (PWID, incarcerated, or homeless individuals) are in marginalized populations that may not have regular access to medical care. As a result, these groups might be missed by existing HCV infection outreach programs that assume some level of healthcare access. The information provided here can help guide outreach to currently HCV-infected individuals, as well as help to identify populations at risk for reinfection or future infection with HCV. In addition, this review was intended to be a starting point for identifying novel risk factors and populations in which HCV infection is underdiagnosed. By summarizing the known information on HCV prevalence, we hope to create a jumping off point to look for as-yet unidentified risk factors or to look for these risk factors in previously uninvestigated populations – whether this expansion is into different social or economic groups, or into different geographic regions.

We also acknowledge several limitations in this umbrella review. The SLRs in this umbrella review included several pooled estimates of HCV infection prevalence that had significant heterogeneity. The underlying studies often included diverse populations and study methods, so the pooled measures should be interpreted with caution. In addition, many of the reviews include studies with cross-sectional designs (see **Table 2.1**), limiting the ability to make inference about causality.

Reviewing only SLRs resulted in gaps concerning possible risk factors that did not (or not yet) have a systematic review. This limitation is particularly relevant in the case of newly discovered or emerging risk factors that may not have sufficient evidence to be reviewed and summarized. For some HCV infection risk factors, such as piercing and tattooing, results from published studies were not conclusive or were limited to specific contexts (for example, in prison, using equipment that may not have been adequately disinfected) [92-94]. For other activities, such as hemodialysis and vertical (i.e. mother-to-child) transmission, new evidence has recently been published[95, 96] and SLRs have not yet been conducted to summarize these findings. In many cases, comorbid conditions and behaviors in individuals with these risk factors for HCV infection make it difficult to assess whether we have identified an independent risk factor for HCV infection or simply a proxy for other known risk factors. Also missing were SLRs quantifying HCV infection prevalence in the broader HIV-positive community. SLRs of US data focused exclusively on HCV infection prevalence in HIV-positive MSM or in HIV-positive MSM who use drugs, making it difficult to understand the contribution of HIV-infection as an independent risk factor for HCV infection.

It was difficult to analyze temporal trends in risk factors, due to the wide range of data collection periods in the underlying studies of each SLR. When multiple reviews for a risk factor or region contained the same studies, the most comprehensive and contemporary review was retained, which, by design, skewed our findings to more current data but obscured insight into trends over time in risk.

We limited our review to prevalence estimates for HCV antibodies. Prevalence estimates for current HCV infection are fewer in number, as they require follow-up testing for HCV RNA. As a result, included prevalence estimates describe individuals with both current and past HCV infection, as opposed to describing only those with current infection.

Next Steps:
Through this review, evidence gaps were discovered that warrant additional consideration. No systematic reviews of HCV infection prevalence associated with hemodialysis met our inclusion criteria, though individual studies have been published. In addition, no SLR was identified that examined vertical transmission of HCV infection in the US. As a result of the opioid epidemic, women of childbearing age are more likely to be infected with HCV than in the past. Finally, there were no systematic reviews of US data characterizing HCV infection prevalence in individuals with tattoos and piercings. While these behaviors were sometimes referenced as potential HCV infection risk factors in reviews examining HCV prevalence among incarcerated persons, they were not the focus of the studies.

By describing the known HCV infection risk factors and identifying the communities most affected by HCV, we have developed a foundation on which further investigation can be initiated. Supervised machine-learning methods can be applied to population-level datasets to build predictive algorithms using these known HCV infection risk factors.

2.6: Conclusion
HCV infection occurs through parenteral exposure to infected blood. This review describes eight risk factors for this exposure in the US. To achieve the WHO 2030 viral hepatitis elimination goal, it is necessary to identify and treat individuals with existing HCV infection and minimize the incidence of future HCV infection through treatment, education, and preventative measures. Understanding the risk factors for HCV infection and the populations most likely to be exposed to those risks will aid in finding and diagnosing HCV-infected individuals and provide a starting point for the discovery of novel risk factors for HCV infection.

2.7: Figures and Tables:

Figure 2.1: Flowchart of search strategy and article inclusion and exclusion criteria

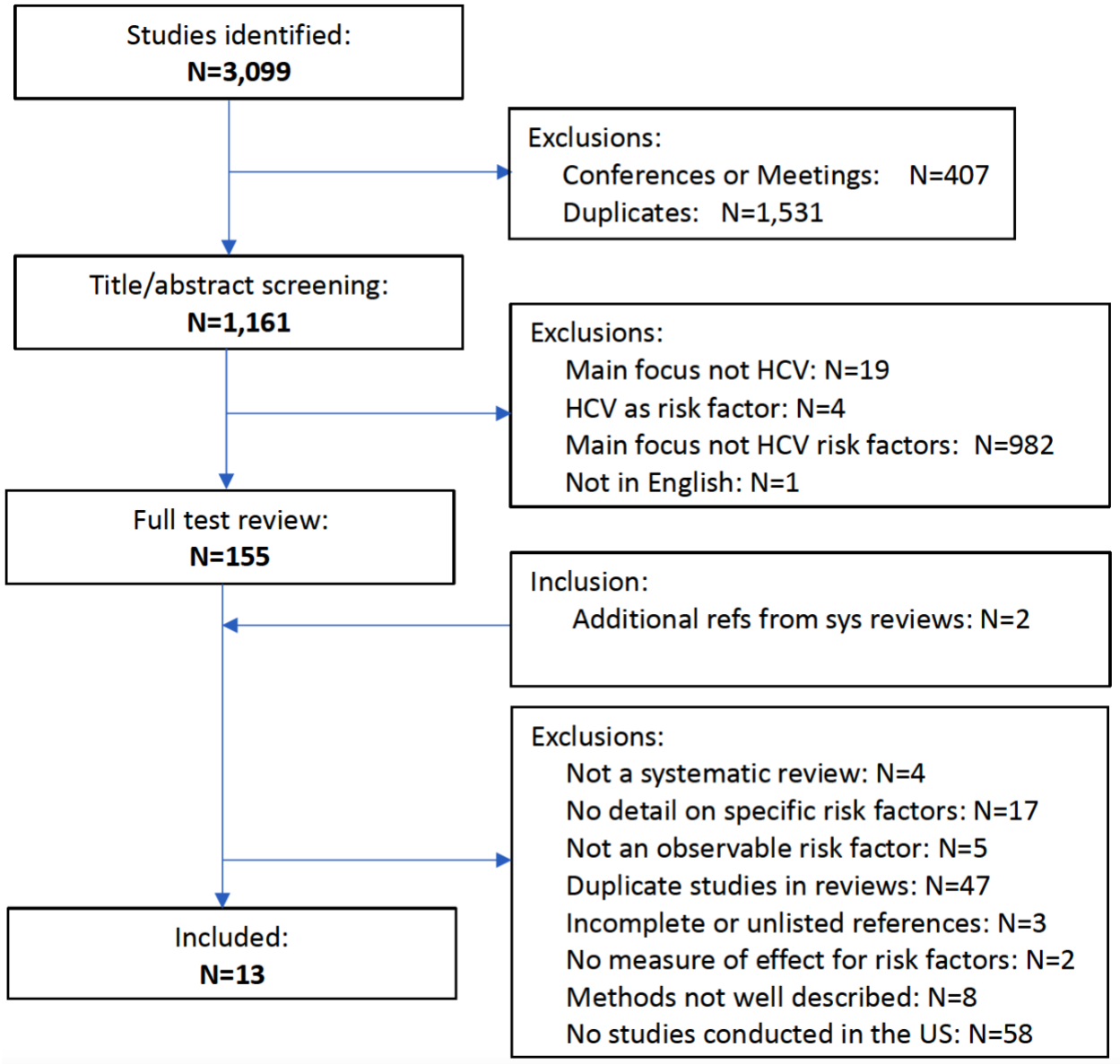


Table 2.1: Study characteristics of systematic reviews organized by HCV infection risk factor

Risk Factor	Author	Year Published	Study Design	Studies (n)	Individuals (n)	Pooled Prevalence	Range	
							Low	High
PWID	Larney S[65]	2013	Cohort studies	3	764		55.5	82.8
	Jordan AE[66]	2017	NR	7	812	35.6 (21.1 – 50.1)	17.9	65.6
	Wirtz AL[67]	2018	NR	6	NR		35.8	84.9
	Bruce V[68]	2019	Cross-sectional, screening, cohort, record review	4	399		25.7	67.6
	Zhou B[69]	2019	Cross-sectional, cohort, case control	2	819		48.3	82.2
	Schalkoff CA[70]	2020	Cross-sectional	3	6,978		34	54.8
	Arum C[71]	2021	Cross-sectional, longitudinal	2	3,581		27.4	82.4
	Jin F[72]	2021	Cross-sectional, cohort	8	NR	29.9 (16.5 – 45.2)		
Non-IDU	Jordan AE[66]	2017	NR	6	3,187	7.5 (5.2 – 9.9)	4.9	12.3
MSM	Wirtz AL[67]	2018	NR	1	NR			50
	Jin F[72]	2021	Cross-sectional, cohort	21	21,523	4.4 (2.9 – 6.1)		
HIV Infection	Jordan AE[66]	2017	NR	7	4,074	12.0 (8.5 – 15.4)	6	17.9
	Jin F[72]	2021	Cross-sectional, cohort	12	11,502	9.7 (7.1 – 12.6)		
Severe Mental Illness	Hughes E[77]	2016	Cross-sectional	12	4,977	17.4 (13.2 – 22.6)	8.2	38
	Ayano G[75]	2018	Cross-sectional, case control	2	5,087		5.9	16.1
HCW	Westermann, C	2015	NR	8	7,213		0	1.9
Incarcerated	Larney S[65]	2013	Cohort studies	9	19,223	29 (24 – 35)	0.8	41.1
	Dolan K[80]	2016	Mostly cross-sectional	20	2,255,000	15.3 (13.1 – 17.7)	20.7	40.5
	Wirtz AL[67]	2018	NR	6	NR		0.1	17
Homeless	Beijer U[81]	2012	Cross-sectional	12	5,391			

Abbreviations: PWID, people who inject drugs; Non-IDU, non-injection drug users; MSM, men who have sex with men; HIV, human immunodeficiency virus; HCW, healthcare workers; NR, not reported.

Table 2.2: Counts of systematic literature reviews, studies, and individuals for each HCV infection risk factor, by risk category

Risk Factor	SLRs (n)	Studies (n)	Individuals (n)
Behavioral/Lifestyle Factors			
PWID	8	35	6,978*
Non-IDU	1	6	3,187
MSM	2	22	29,523*
Medical Conditions			
HIV Infection	2	19	15,576
Severe Mental Illness	2	14	10,064
Occupations			
Healthcare worker	1	8	7,213
Special Populations			
Incarcerated	3	35	2,274,223*
Homeless	1	5	1,758

* Underestimated - systematic literature review did not report enrollment for individual studies
Abbreviations: PWID, people who inject drugs; Non-ISU, non-injection drug users; MSM, men who have sex with men; HIV, human immunodeficiency virus;

2.8: Supplemental Material

Supplemental Table S2.1: Systematic literature review search terms

Search Terms	("hepatitis c" or "HCV" or "hep c" or HEPC).ti.ab AND (systematic review OR systematic literature review OR systematic scoping review OR systematic narrative review OR systematic qualitative review OR systematic evidence review OR systematic quantitative review OR systematic meta-review OR systematic critical review OR systematic mixed studies review OR systematic mapping review OR systematic cochrane review OR systematic search and review OR systematic integrative review).ti OR systematic review.pt NOT (conference or meeting).pt

Supplemental Table S2.2: Resources searched on December 5, 2021

Name	Description and Date Range
BIOSIS Previews	1993 to 2021 Week 52
Embase	1974 to 2021 December 03
EBM Reviews	Cochrane Central Register of Controlled Trials: November 2021
EBM Reviews	Cochrane Database of Systematic Reviews: 2005 to December 02, 2021
EBM Reviews	ACP Journal Club: 1991 to November 2021
EBM Reviews	Cochrane Clinical Answers: November 2021
EBM Reviews	Database of Abstracts of Reviews of Effects: 1st Quarter 2016
EBM Reviews	Cochrane Methodology Register: 3rd Quarter 2012
EBM Reviews	Health Technology Assessment: 4th Quarter 2016
EBM Reviews	NHS Economic Evaluation Database <1st Quarter 2016>
Ovid MEDLINE®	Ovid MEDLINE® and Epub Ahead of Print, In-Process, In-Data-Review and Other Non-Indexed Citations, Daily and Versions®: 1946 to December 3, 2021

Supplemental Table S2.3: Summary characteristics of included systematic literature reviews, listed alphabetically by author

Author	Pub Year	Data Start	Data End	Risk Factors	Study Designs	Studies (n)	Participants (n)
Arum C[71]	2021	2000	2017	PWID	Cross-sectional, longitudinal	2	3,581
Ayano G[75]	2018	1994	2017	Severe mental Illness	Cross-sectional, case control	2	5,087
Beijer U[81]	2012	unrestricted	2012	Homeless	Cross-sectional	5	1,758
Bruce V[68]	2019	1995	2017	PWID	Cross sectional, record review, screening, cohorts	4	399
Dolan K[80]	2016	2005	2015	Incarcerated	Mostly cross-sectional	20	2,255,000
Hughes E[77]	2016	unrestricted	2015	Severe mental Illness	Cross-sectional	12	4,977
Jin F[72]	2021	2000	2019	HIV-infection, MSM, PWID	Cross-sectional, cohort	21	29,523
Jordan AE[66]	2017	1990	2015	HIV-infection, non-IDU, PWID	NR	7	4,074
Larney S[65]	2013	unrestricted	2012	Incarcerated, PWID	Cohort studies	9	19,223
Schalkoff CA[70]	2020	2006	2017	PWID	Cross-sectional	6	1,979
Westermann C[79]	2015	1989	2014	HCW	Cohort studies	8	7,213
Wirtz AL[67]	2018	2005	2015	Incarcerated, MSM, PWID	NR	6	NR
Zhou B[69]	2019	2005	2019	PWID	Cross-sectional, cohort, case control	3	5,286

Abbreviations: PWID, people who inject drugs; Non-IDU, non-injection drug users; MSM, men who have sex with men; HIV, human immunodeficiency virus; HCW, healthcare workers; NR, not reported.

Chapter 3: Identifying risk factors associated with hepatitis C virus infection in participants in the National Health and Nutrition Examination Survey using Super Learner

3.1: Abstract

Under-diagnosis is a key impediment to eliminating hepatitis C virus (HCV) infection in the United States (US). To address this problem, the United States Preventative Services Task Force (USPSTF) updated HCV screening guidelines in 2020 to include one-time testing for HCV antibodies in adults (18 – 79 years of age). Machine learning methods offer an opportunity to support increased screening in a resource-optimized way. Directing screening resources towards those most likely to be infected with HCV will reduce screening costs by minimizing testing of uninfected people, while minimizing the burden of unnecessary testing on those not infected.

In the current study, the ensemble machine-learning method Super Learner was used with National Health and Nutrition Examination Survey (NHANES) data collected from 2013 – 2016 to build an HCV infection prediction algorithm. Because only 1.1% of survey participants tested positive for HCV RNA, case control sampling was used to address class imbalance, with five HCV-uninfected individuals randomly selected for each HCV-infected individual. Performance of the predictor was evaluated by calculating the area under the precision recall curve (AUC-PR) and comparing precision at different levels of recall to the level of precision that would occur with perfect uptake of universal screening. The final fitted algorithm was used to predict HCV infection in the full NHANES dataset and to identify the features in the data that had the greatest impact on the AUC-PR, as they represent the characteristics, were most predictive of HCV infection.

The fitted Super Learner produced a final AUC-PR of 52.0% on the full dataset. By choosing a probability threshold that optimized the F1 score, the algorithm achieved 55.4% precision and 61.3% recall, meaning that for every 100 individuals classified as HCV-infected by the algorithm, more than 50% would be true positives if the goal was to identify at least 60% of those infected with HCV. Features that were most predictive of HCV infection included alanine aminotransferase (ALT), aspartate aminotransferase (AST), injection drug use, age, albumin, globulin, and HBV infection status. Additional influential predictors included smoking-related features, features indicating risk factors for heart disease or stroke (triglycerides, total cholesterol, recommendations by a doctor to take low dose aspirin), and features related to oral health.

As outreach continues to identify the undiagnosed HCV-infected population in the US, knowledge of the characteristics most predictive of HCV infection should be utilized to guide screening and prevention efforts. Algorithms such as the one demonstrated here can support these activities.

3.2: Introduction

Hepatitis C virus (HCV) is single stranded positive sense RNA virus that preferentially infects hepatocytes, resulting in significant morbidity and mortality globally[5]. HCV infection is characterized by liver inflammation and scarring and can progress to cirrhosis, liver

decompensation, and hepatocellular carcinoma (HCC). It is estimated that 1.7% (95% confidence interval; CI 1.4 – 2.0%) of the United States (US) population are HCV antibody positive, and 1% (95% CI: 0.8 – 1.1) have active viremia[57], making HCV one of the most common bloodborne viruses in this population[97, 98]. Known risk factors for HCV infection include exposure to infected blood through injection drug use, blood transfusion prior to HCV screening of blood products (pre-1992), piercing and tattooing with unsterilized needles, and occupational exposure, such as in healthcare work.

While 15-45% of those infected with HCV will spontaneously clear the virus, most infections become chronic and require treatment[10, 99, 100]. The approval of direct acting antivirals (DAAs) to treat HCV infection profoundly simplified treatment and improved outcomes for people infected with the virus[15, 16]. Early treatment not only prevents complications associated with chronic infection[101], but also breaks the cycle of transmission. The World Health Organization (WHO) has challenged health agencies around the world to eliminate viral hepatitis (B and C) by reducing new infections by 90% by 2030[17]. A key impediment to eliminating HCV infection is lack of diagnosis, as the majority of infections go undetected, and most HCV-infected individuals are symptom free until significant liver damage has occurred. In fact, 80% of people with HCV infection globally, and 50% of those infected in the US remain unaware of the infection[18, 19, 102]. The US has employed a variety of strategies to improve the diagnosis of HCV infection, including a change in screening recommendations to universal one-time screening for all adults 18-80 years of age, with more frequent screening for pregnant women and those with known risk factors for infection[34, 103]. However, in the face of limited resources, stigma associated with HCV infection and the competing demands of the global Covid-19 pandemic, it's unclear whether uptake of the new recommendations has been successful. Historically, recommendations to increase HCV screening in the US have been slow to gain traction[35].

Machine learning methods offer an opportunity to support increased screening in a resource-optimized way. Large, richly-featured datasets, such as administrative claims data or electronic medical records, can be analyzed to identify complex relationships between known and unknown risk factors to help identify people with a high probability of infection with HCV and prioritize them for HCV screening. Applying screening resources to those most like to be infected with HCV will reduce screening costs by minimizing testing of uninfected people, while minimizing the burden of unnecessary testing on those not infected[104]. From these models, a list of highly predictive characteristics of HCV infection can be identified that could be used by public health organizations to guide HCV infection prevention and education and increase diagosis of undiagnosed HCV-infected individuals.

In machine learning models, a variety of different algorithms can be applied, individually and in combination (ensemble learning), to form predictions. In addition, supervised learning models (models built on datasets where the outcome is already known) are assessed in a cross-validated way by comparing prediction of an outcome to actual outcome status on data that were not included when building the algorithm, to maximize signal detection and avoid overfitting to random variation in the data. These models can then be applied to novel data with the same features that went into creating the prediction algorithm, to produce a probability of the outcome of interest.

There is increasing interest in using machine learning to predict various outcomes in patient data. For example, Dinh *et al.* used data from the National Health and Nutrition Examination Survey (NHANES) to develop models that identify patients with cardiovascular disease and with diabetes mellitus, and to identify the strongest predictors of these conditions in the dataset. They achieved an area under the receiver operator characteristic curve (AUROC) of 83.9% for cardiovascular disease using an ensemble model (a model that combines multiple predictive algorithms), and 95.7% for diabetes mellitus using an eXtreme Gradient Boost model. They also identified the top five predictors in that dataset for each condition[105]. Oh *et al.* also used NHANES data from 1999 – 2012 to create a deep-learning algorithm to predict depression in Korean NHANES (K-NHANES) data from the same time period and in NHANES data from a different time period (2013-2014). Their algorithm had prediction success cross-culturally, with an AUROC of 0.77 on the K-NHANES data, and cross-temporally with an AUROC of 0.92 on the NHANES data from 2013-2014[106]. Doyle et al. used US administrative claims data with a stacked ensemble model to predict HCV infection. The resulting algorithm achieved 97% precision (PPV) at 50% recall (sensitivity), which can be interpreted to mean that for every 100 individuals flagged as HCV-positive by this algorithm, approximately 97 would be true positives and 50% of individuals with HCV infection would be identified. At that time, the Centers for Disease Control and Prevention (CDC) recommended HCV screening for all persons born between 1945 and 1965, in whom the prevalence of HCV infection was estimated to be 2.2%; thus, the algorithm would have significantly outperformed the CDC screening guidelines in identifying HCV-infected individuals[104]. Given that HCV infection is both rare and significantly underdiagnosed, these examples highlight the potential benefit of using machine learning as a tool to improve identification of those who were HCV-infected.

In the current study, the ensemble machine-learning method Super Learner was used with 10-fold cross validation to predict HCV infection in NHANES data collected from 2013 - 2016. Performance was evaluated by calculating the area under the precision recall curve (AUC-PR) and comparing precision at different levels of recall to precision that would occur with perfect uptake of universal screening. From the resulting algorithm, those features in the NHANES dataset that have the greatest impact on the AUC-PR were identified, as they represent the characteristics that are most influential in HCV infection prediction.

3.3: Methods

Data

The primary goal of this study was to identify the risk factors most predictive of HCV infection. Because many people who are infected with HCV but have not yet progressed to severe liver disease and are not experiencing related signs and symptoms, we wanted to use data that didn't rely on symptom-driven interaction with a healthcare provider. People who are uninsured or under-insured will likely be under-represented in administrative healthcare data; however, they are an important population to include in such research, as they may have higher rates of chronic HCV infection than those who have fewer barriers to medical care. Finally, a supervised machine learning approach was employed to develop the HCV prediction algorithm, which necessitates working with labeled data (data that include a measurement for the outcome of interest: the presence of HCV RNA). NHANES data meet all of the above requirements and were, therefore,

selected for this research. These data are collected from a random sample of US residents; selection into the survey does not require health insurance, and all participants who complete the examination portion of the survey are tested for HCV RNA.

NHANES is an ongoing population-based cross-sectional study conducted by the National Center for Health Statistics (NCHS), a branch of the CDC. The study began in 1960 to comply with the National Health Survey Act, which was passed by the US Congress in 1956 with a mandate to characterize the distribution of illness and disability in the United States. In order to be representative of the civilian non-institutionalized population, survey participants are selected using a complex multi-stage probability sampling design. Populations of particular interest are oversampled to increase the precision of subgroup estimates; therefore, survey weights are included with the data. Sectors of the population not included in the survey include persons who are incarcerated, homeless, in nursing homes, on active military duty, or living outside of the 50 states and the District of Columbia.

NHANES data are captured through both a home interview and a physical examination in a mobile examination center (MEC). The data collection study team includes physicians, medical technicians, and trained health interviewers who speak English and Spanish. An initial doorstep interview is conducted to ascertain whether anyone in the selected residence is eligible to participate in the study and the relationships between all individuals living there. Once a participant is identified, the home interview is conducted, including questions about demographic factors, socioeconomic status, diet, and disease history. A follow-up visit is then conducted at the MEC, where a full physical exam is performed, blood and urine samples are obtained, and additional interview questions are asked. Transportation is provided to and from the MEC, if needed, to maximize participation of the randomly selected study sample. In the three most recent two-year cycles, the response rates for completing both components of the survey were 68.5% (2013-2014), 58.7% (2015-2016), and 48.8% (2017-2018), resulting in approximately 9,000 respondents per cycle.

Participant Inclusion/Exclusion Criteria

In this study, people were excluded for any of the following three reasons:

1. Did not participate in the examination portion of the survey
2. Under 20 years of age,
3. No information on HCV RNA.

Note that these groups are not mutually exclusive. We focused on the adult population (i.e. 20 years of age and older) because many questions and laboratory values in the NHANES dataset are not captured for individuals younger than 20 years of age. In addition, because we employed supervised machine learning models that require labeled training data (i.e. knowledge of HCV RNA status), so individuals who did not participate in the examination section of the survey, and who therefore were not tested for HCV RNA, were not included.

Data Cleaning and Feature Engineering

Data from the three most recent cycles of NHANES (2013-14, 2015-16, 2017-18) were combined to form the initial cohort for this analysis. Prior to 2013, CDC guidelines recommended that people who screened positive for HCV antibodies be given a confirmatory recombinant immunoblot assay (RIBA) test before they were tested for current HCV infection

(the presence of HCV RNA) because only about 2/3 of the people who were reactive in the first screening were RIBA positive. However, the company that made the RIBA kits discontinued them at the end of 2012. As a result, guidelines were changed starting in 2013 – 2014, such that people who screen positive for HCV antibodies are tested directly for HCV RNA without a second confirmatory test[107]. Because this change would affect who is tested for HCV RNA, it was decided not to include survey cycles prior to 2013 with those after 2013. In addition, treatment changes that occurred at the end of 2013, coupled with the demographic shift brought about by the opioid epidemic, likely influenced the risk factor profile for current HCV infection; as a result, this analysis focused on the more recent NHANES population.

NHANES data are grouped into five categories: Demographic variables, Examination, Laboratory, Questionnaire, and Diet, with over 4,000 potential datapoints collected for each participant. An additional category of limited access data was not considered for inclusion in this study. Initially, all available raw data were extracted from the NHANES website from all categories available from 2013 - 2018 for pre-processing. This phase involved a variety of steps, beginning with the elimination of covariates that were not consistently available across the three cycles being analyzed. A decision was made to omit diet data, as these data lean heavily on patient reported information that can introduce subjectivity and noise into the machine learning process and have a significant impact on performance. In addition, diet is unlikely to be captured in many publicly available data sets, making it hard to operationalize any diet-related findings. Among the remaining covariates, administrative questions and duplicate laboratory values with different units were excluded. In laboratory data, the international system of units (SI units) was preferentially selected when results were offered in multiple types of units. Some features were aggregated when, for example, the same question asked of different age groups was treated as distinct questions. A composite dental score was created, in consultation with a dentist, to address individual tooth assessments that resulted in hundreds of features that would likely lack power individually but might be meaningful in the aggregate.

Each covariate available in NHANES is a potential feature for the machine learning prediction algorithm. If a feature included a “Refused to answer” (7-series) or “Don’t know” (9-series) option, these were recoded to “NA” to reflect the absence of a definitive answer (imputation of missing values will be described later in this section). Features such as numeric laboratory results were retained as numeric values, and features that included categorical responses were recoded as factor variables in R. Binary Yes/No features were retained wherever reasonable and recoded as 1/0/NA. Once patient inclusion/exclusion criteria had been applied and the full final cohort (not the case-control sample described below) was defined, any feature with a missing response from more than 50% of participants was excluded. For example, some features had a high rate of missingness due to skip logic (i.e. questions only answered based on the conditional response of a previous question). Exceptions to this approach were as follows: (1) all participants were required to have a non-NA value for HCV RNA (and no imputing was done) and (2) because no prescription drug was taken by >= 50% of the participants in the month preceding the interview portion of the survey (the criterion used by NHANES to record medication use), prescription drugs were grouped by drug category and patients were flagged (1/0) if any prescription from a particular category was recorded during the month prior to the survey. NHANES uses Lexicon Plus®, a database owned by Cerner Multum, Inc., which includes all drug products available with and without prescription in the US.

Case-Control Sampling:
The Viral Hepatitis Surveillance Report from the CDC estimates the prevalence of HCV infection to be approximately 1% in the US[4]; therefore, it was expected that the presence of HCV RNA among participants of NHANES would be rare. Many machine learning algorithms assume classes are balanced (i.e. that there are roughly an equal number of people with and without the outcome); therefore, it is likely that a prediction algorithm that is trained on the full dataset would perform poorly. To address this challenge to successful prediction, a case-control sampling approach was employed to allow the model to be trained in an environment with a greater concentration of cases. All participants with HCV infection (i.e. cases) were included in the sampled data set, and participants without HCV infection were randomly selected in a 5:1 ratio to those with HCV infection.

Modeling Process
The first step in building a model with Super Learner (SL) is to define the machine learning task. Within the HCV prediction task, the following information was specified:

- Data set: the NHANES case-control sample
- Outcome variable: HCV RNA
- Covariate features: all covariates except HCV RNA in the data set
- Number of folds for cross-validation (10, stratified by outcome class to ensure balance)
- Weights: no weights were used to train the model

When the task is created, Super Learner imputes missing values for covariate features. For continuous covariates, the median is imputed; for binary and categorical covariates, the mode is imputed. For each covariate, if any imputation is required, an additional covariate is created to flag the values that were imputed as a way to detect whether there are meaningful patterns in the missingness. For example, if every covariate had some values missing, then the feature list would double.

The rare outcome and large number of potential predictors in the dataset present significant challenges to successful prediction of HCV infection. Therefore, an ensemble learning approach was applied to approach this problem from multiple perspectives. Super Learner is a loss-based ensemble learning method that uses v-fold cross validation to create a meta-model from a weighted combination of algorithmic approaches that optimize the specified loss or evaluation function. A diverse library of parametric and non-parametric algorithms was selected to maximize the predictive power of our final model. Parametric learners are faster, require less information, and perform well with simple prediction problems, whereas the non-parametric learners have greater flexibility to identify complex relationships and don't fall prey to misspecification:

Parametric Learners:

- Generalized linear models (*glm*): GLM fits a standard generalize linear model
- *bayesglm*: Bayesglm is an alternative to GLM that uses student-t prior distributions for the coefficients to produce more stable estimates. In a model with many features, especially those with low variance, the use of priors can significantly reduce efficiency. Bayesglm uses a modified expectation-maximization algorithm to fit the model[108, 109].

- *glmnet*: Glmnet estimates a regularized generalized linear model. Penalty options include the range from l1 (Least Absolute Shrinkage and Selection Operator; *LASSO*), which can shrink the slope of unhelpful coefficients to 0, effectively removing them from the model, to l2 (ridge) which keeps all features but shrinks their slopes to close to 0, and mixtures of LASSO and ridge (elastic net). The mixture is established by a specified alpha (α) that ranges from 0 (ridge) to 1 (LASSO)[110]. We used glmnet learners with alphas ranging from: α= [0, 0.2, 0.4, 0.6, 0.8, 1].

Non-Parametric Learners:

- Extreme gradient boosting (*xgboost*)[111]: Xgboost uses gradient boosted decision trees and is optimized for speed and model performance. Boosting builds models in a sequential manner and uses a loss function and weights to focus on the most challenging cases (sequentially higher weights are given to misclassifications for subsequent iterations). The number of fitting iterations was set to 500, and early stopping rounds was set to 50. Early stopping, a way to avoid over-fitting,is used when the loss on the validation set starts to increase.
- Discrete bayesian additive regression tree sample (*dbarts*)[112]: Dbarts is a Bayesian tree ensemble method that uses individual trees as base learners. Each tree is constrained by a prior to be a weak learner. This learner is flexible and requires minimal assumptions.
- *Ranger*: Random forest (an ensemble method using decision trees and bagging) is optimized for high dimensional data.[113] Ranger builds on random forest by allowing the user to choose a mode for calculating variable importance, the contribution a specific feature makes to prediction. Importance criteria "impurity" was used (impurity measurement uses the Gini index[114] for classification that is the probability that a randomly selected feature is classified incorrectly).

The latest version of Super Learner (SL3) also includes the ability to create pipelines where one can add additional screening options. Pipelines perform functions sequentially, using the result of the first function in the second, and so on. Pipelines were used in two ways in our learning library.

- First, they were used to screen covariates with a penalized regression method (LASSO) prior to being fit with the learners *glm* and *bayesglm*. This was done becasue *glm* and *bayesglm* both attempt to use every covariate that would lead to significant overfitting, given our sample size.
- Next, a second pipeline was created using the learner *ranger* to identify the 100 most influential covariates (based on the selected evaluation metric – see below). These selected covariates were then fed to the full stack of learners in the library (including the learners from the first pipeline, *glm* and *bayesglm*). Screening was employed to exclude features that didn't contribute meaningfully to prediction of HCV infection. The remaining features should be those that will perform best on out-of-sample prediction.

After the individual base learners were defined, a metalearner was specified with the job of creating a weighted combination of the predictions from the individual learners. We used the default metalearner, nonlinear optimization via augmented lagrange.

Once the machine learning task, learners, and screening pipelines were defined, the following steps were executed:

1. Super Learner with 10-fold cross validation was used to build a model to predict HCV infection in the sample data
2. Cross-validation of the Super Learner fit (again using 10-folds) was performed to assess performance on unseen data.
3. The fitted Super Learner was used to predict HCV infection on the full NHANES dataset, with weights to represent the US population surveyed during the six-year period.
4. Variable importance was calculated. Characteristics that had the largest impact on prediction of HCV infection were identified (described below).

Performance metrics

In imbalanced classification problems (where most individual in the dataset do not experience the outcome of interest – called negative cases), assessing prediction success with the area under the receiver operation characteristic curve (AUROC) is not an ideal metric. The AUROC compares the proportion of those with the outcome who test positive. [(sensitivity) $\frac{TP}{(TP+FN)}$] to those without the outcome who test positive [(1-specificity); $\frac{FP}{(FP+TN)}$]. Given that HCV infection is present in only 1.1% of the full NHANES set of participants (see above), predicting each case as negative will produce almost 99% accuracy, yielding a strong AUC without producing successful prediction. Even with case control sampling, basing performance on the AUROC would yield overly optimistic results.

Area under the precision recall curve (AUC-PR) is preferable to the AUROC when the outcome of interest is rare. Precision-recall curves focus on the positive class identification (even if it's rare) and don't factor in the success of predicting true negatives. The geometric interpretation of the precision recall curve is the expected precision: $\frac{TP}{(TP+FP)}$ when uniformly varying the recall: $\frac{TP}{(TP+FN)}$. Flach and Kull explained the limitations of precision-recall curves (lack of universal baseline; uninterpretable region on the lower right side of the graph; lack of calibration; etc.)[115]. Generally, precision and recall can be assessed through the F-score (F combines precision and recall into one metric, and is more useful than accuracy when you have class imbalance), where, when $\beta = 1$, and F is the harmonic mean of precision and recall[116]. The value of β determines the tradeoff between precision and recall. When $\beta > 1$, recall is given greater weight, when $\beta < 1$, precision is given more weight.

$$Accuracy = \frac{True}{True+False} = \frac{TP+TN}{TP+TN+FP+FN} \qquad \text{where as} \quad F_\beta = (1 + \beta^2) * \frac{Precision*Recall}{(\beta^2*Precision)+Recall}$$

The baseline in a precision recall curve is equivalent to the true prevalence of the outcome at all values of recall (forming a horizontal line). At maximum sensitivity, all samples are predicted to be cases.

Variable importance

As a final step, variable importance was calculated by considering AUC-PR with and without each feature, to see which features had the largest impact on the metric. The features with the largest impact would be useful for identifying expanding the search of people with current HCV infection. We used the option to "permute" rather than remove the covariate for computational efficiency. Both a difference and a ratio measure were calculated, where:

- Difference = (AUC-PR without feature) – (AUC-PR with feature).
    - Thus, larger negative values are associated with features that have a higher impact on AUC-PR
- Ratio = (AUC-PR without feature) / (AUC-PR with feature)
    - Thus, smaller fractional values are associated with features that have a higher impact on AUC-PR

3.4: Results:

Descriptive Results

The three NHANES survey cycles from 2013-2018 included 29,400 participants from 90 unique primary sampling units (PSUs), where each PSU represented a single large US county or smaller contiguous counties. Among those selected for the survey, 1,339 (4.8%) did not participate in the examination. An additional 11,734 (41.8%) of the remaining 28,061 were under 20 years of age. Finally, 1,090 (6.7%) of the remaining 16,327 did not have screening results for HCV RNA. After these inclusion/exclusion criteria were applied, 15,237 people remained in the full NHANES HCV prediction cohort.

Of the 15,237 individuals in the full cohort, 168 participants tested positive for HCV RNA, resulting in a prevalence of 1.1% (before applying weights). From the 15,069 participants with no evidence of HCV RNA, five negative individuals were randomly selected for every HCV-infected case, resulting in 840 controls, and an effective sample size of 1,008 survey participants in the training dataset (**Figure 3.1**). No weights were included when training the prediction algorithm, so as to maintain the 5:1 non-case to case ratio; however, weights were added when prediction was done on the full 15,237-person dataset to assess overall performance of the algorithm.

Key characteristics of the training cohort are provided in **Table 3.1**. Mean age was slightly higher in the HCV positive cohort (56.7 years, (standard deviation; SD: 10.8) vs. 50.8 years (17.9)). People with HCV were more likely to be male (70.8% for HCV infected vs. 48.3% for HCV uninfected), living below the poverty level (49.4% vs. 21.6%), and have a history of injection drug use (46.5% vs. 2.3%), and less likely to be married (38.7% vs. 58.5), college-educated (32.7% vs 56.4%), and obese (30.1% vs. 42.5%). Inclusion/exclusion criteria were also applied to the full set of potential HCV predictors. From the 4,134 covariates in the raw data, the preprocessing stage identified 1,099 (excluding prescriptions medications, which were handled separately) (see **Table 3.2**).

As a next step, any feature that was not available for at least 50% of the 15,237 participants in the full cohort was excluded, resulting in 369 remaining covariates. After adding in indicator variables for the 17 prescription drug categories, a total of 386 potential predictors of HCV infection were identified. When the machine learning task was defined for Super Learner, 303 features in the dataset had missing values that required imputation, resulting in the creation of

303 additional covariates to assess patterns of missingness. In total, 689 features were used by Super Learner for prediction of HCV infection. (**Figure 3.2**).

Prediction Results

Models for each individual algorithm described in the methods section were developed using the 369 individual features, 17 prescription categories, and 303 imputed covariate flags. The Super Learner divided the 1,008 participants into 10 folds (with HCV-infected participants evenly distributed in each fold). Nine folds were used to build a model to predict HCV infection and then that model was tested on the left-out 10^{th} fold. This process was repeated for each permutation of nine folds, with validation on the 10^{th} fold, until a matrix was built with out-of-sample predictions for every survey participant on every algorithm in the learner library. This matrix was then used to build an ensemble learner that consisted of a weighted convex combination of the predictions of the individual learners. (**Figure 3.3**)

Prediction success was assessed by the ability to maximize AUC-PR. Again, the baseline for AUC-PR would be the prevalence of HCV in the sample, in this case 16.7%. Results from the individual learners and the ensemble learner are provided in **Table 3.3**. Non-parametric learners (ranger, dbarts, xgboost) generally outperformed parametric learners. Among the top four learners, three were decision-tree-based algorithms. AUC-PRs from the individual learners were very similar, ranging from 87.2% – 89. 9%. The fitted Super Learner achieved an AUC-PR of 90.1%, exceeding the individual models. Each of the 11 learners was weighted equally at 0.0909 to create the ensemble learner. Reasons for the identical contribution of each learner are unclear; however, it may be related to fact that individual learners performed similarly.

The fitted Super Learner was then cross validated so that performance of the ensemble learner could be assessed on unseen data. This process adds an additional layer of cross validation to the Super Learner process, dividing the data into 10 folds, setting aside 10% of the data (a validation set), and then dividing the remaining 90% of the data into 10 folds and repeating the process described above to create a super learner. The resulting fitted super learner was assessed on 10% validation set. This process was repeated 10 times, leaving out a different 10% validation set each time, building a super learner on the remaining 90% of the data, then validating on the hold out sample. Results from the cross-validation of the super learner can also be found in **Table 3.3.** Note that XGBoost, ranger, and dbarts, the three non-parametric learners, performed best in the cross-validated super learner, all with an AUC-PR over 89%. The super learner outperformed individual learners, achieving an AUC-PR of 90.1% on unseen data.

Once the super learner was built and cross-validated, it was used to predict a probability of HCV infection for each individual, first on the NHANES sample, (N = 1,008), then on the full dataset (N = 15,237). Additional performance metrics are presented in **Table 3.4**. In the NHANES sample, the final AUC-PR was 97.9%. To calculate recall (sensitivity) and precision (positive predictive value), a prediction threshold was chosen to optimize the F1 score. In the NHANES sample dataset, where the prevalence of HCV infection was 16.7%, a threshold of 38.2% was selected (a predicted probability of HCV infection >= 38.2% was assumed to be HCV +). This threshold produced 94.4% precision and 98.9 % specificity at a recall level of 90.5%, meaning that, out of every 100 individuals classified as HCV+, 94 would be true positives, if the objective was to find at least 90.5% of HCV-infected individuals.

In the full data set, HCV infection was comparatively rare, with a prevalence of 1.1%. The fitted super learner produced a final AUC-PR of 52.0%. A stricter threshold of 0.854 was needed to optimize the F1 score, as the full dataset has far more true negatives. Precision recall graphs are presented in **Figure 3.4**. At this threshold, precision was 55.4% and specificity was 99.4% at a recall level of 61.3%, meaning that, out of every 100 individuals classified as HCV infected, over half (55) would be true positives, if the goal was to find at least 60% of people infected with HCV. (See Supplemental Figure S3.1 for AUC-PR graphs of individual learners)

Variable importance

Features that had the largest impact on AUC-PR were identified both as a difference and as a ratio. **Figure 3.5** includes the 15 features with the most influence on AUCPR, calculated as a difference (AUCPR without the feature – AUCPR with the feature). **Figure 3.6** shows the 15 features with the greatest influence on AUCPR, calculated as a ratio (AUCPR without feature / AUCPR with feature). Overall, alanine aminotransferase (ALT), aspartate aminotransferase (AST), and injection drug use were identified as the most important predictors of HCV infection by both difference and ratio measures. Age, albumin, globulin, and HBV infection status were also identified by both metrics as features highly predictive of HCV infection. Additional influential predictors included smoking-related features, features indicating risk factors for heart disease or stroke (triglycerides, total cholesterol, recommendations by a doctor to take low dose aspirin), and features related to oral health.

3.5: Discussion

Underdiagnosis of HCV infection is a substantial problem. Up to 80% of people globally and 50% of people in the US who are infected with HCV are unaware of their infection. Untreated, HCV infection can become chronic and lead to severe long term health consequences, as well as being passed onto others through exposure to infected blood. To support the WHO goal of HCV elimination by 2030, the United States Preventative Services Task Force (USPSTF) updated guidelines to recommend one-time universal screening of all adults for HCV infection. While this change will undoubtedly identify additional individuals with HCV infection, it also will result in substantial resource utilization with a low yield, given the estimated 1-4% prevalence of HCV infection in the US adults. In addition, it is unclear whether or how quickly the new universal one-time screening guidance will be embraced.

The goal of this study was to support accelerated identification of HCV-infected individuals by developing an infection prediction algorithm using annual survey data collected from a random sample of the US population and then identifying the characteristics that played the largest role in successful prediction. This list of characteristics could be used by public health agencies for education and community outreach, as well as to create additional avenues for identifying priority candidates for HCV screening. The use of ensemble learning provided the opportunity to predict HCV infection in NHANES participants with multiple robust approaches, both parametric and non-parametric, and to create the best combination of those methods to maximize prediction power.

Previous studies have successfully used machine learning methods with large datasets to advance identification and treatment of HCV infection. Haga et al. used machine learning methods for genomic analysis to identify individuals infected with HCV variants that were resistant to DAAs[117], while Doyle et al. used ensemble learning with administrative claims data to support earlier identification of HCV infected individuals [104]. We used an ensemble learning approach with cross-sectional national survey data to better understand the characteristics associated with HCV infection.

To the best of our knowledge, this is the first study to use ensemble methods with NHANES data to predict HCV infection. No previously published studies of HCV infection have included these data, as the NHANES 2017-2018 results for HCV antibody and RNA testing have only recently become available. This combination of current data and innovative methods provides an opportunity to improve identification of the HCV-infected population in the US. Other strengths of this study include the fact that HCV antibody screening was performed on all NHANES MEC participants, without regard for identified risk factors or symptoms, enabling the prediction algorithm to include characteristics of HCV-infected individuals who may have evaded previous identification efforts. Additionally, no health insurance is required for NHANES participation, enabling the survey to include information about people who may not have consistent access to health care and who might, therefore, be at greater risk of having an undiagnosed HCV infection.

The algorithm produced for this study showed a significant increase in precision in identifying HCV infection across all levels of recall, compared to universal screening, providing evidence that an algorithm produced with machine learning methods could make an important contribution to case identification. The stacked ensemble created by Super Learner achieved over 94% precision at 90% recall in the NHANES sample dataset of 1,008 participants, where universal screening would have achieved 16.7% precision at all levels of recall. In the full data set, the algorithm achieved over 50% precision at 61% recall, where universal screening would have achieved 1.1% precision at all levels of recall. In the full data set, precision begins to decline as recall extends beyond 60% (i.e. at higher levels of sensitivity).

To identify an HCV-infected individual, connect that individual with treatment, and prevent the spread of the infection to others, it might be considered a reasonable tradeoff to have high recall and incur more false positives to maximize identification of true positives. Additionally, in the universal screening scenario, testing everyone reduces some of the stigma that may accompany HCV screening. However, many people find blood tests (which are necessary for HCV screening) stressful and unpleasant, and HCV screening still may not be prioritized in the minimal time allotted to physician-patient encounters in the US, particularly in situations with no obvious symptoms or risk factors present. If an algorithm can identify subtle relationships that go beyond obvious risk factors for HCV infection with a significantly higher degree of precision than would occur through universal screening, it may provide increased motivation to prioritize screening of the individuals identified.

The 15 features in the NHANES dataset most predictive of HCV infection could be grouped into three categories: (1) known risk factors, (2) factors associated with the 1945-1965 birth cohort, and (3) a group of characteristics that don’t fall into the first two categories. The presence of known risk factors, including indicators of Group 1 such as ALT, AST, HBV infection, albumin

(a protein produced by the liver), globulin (abnormal levels can indicate infection), lymphocytes, and injection drug use, inspire confidence that the algorithm is working correctly to incorporate current knowledge to predict HCV infection. Group 2 characteristics, such as age, triglycerides, total cholesterol, and features related to cardiac health (including sagittal abdominal diameter[118]), could represent indicators associated with aging. Because the 1945 – 1965 birth cohort is now 56-76 years of age, it is conceivable that features in this category simply identify people in the birth cohort who are known to be at greater risk of HCV infection due to exposure to infected blood prior to HCV screening of blood products. The third group of highly influential characteristics includes some unexpected characteristics, including cigarette smoking, which could potentially be a proxy for other high-risk behaviors that people are less willing to acknowledge. Additionally, indicators of poor oral health appeared on the list of important features, including tooth count and tooth condition, which were not among the top 15 most influential predictors, but which still had an impact on the AUC-PR. Current research suggests an32ssociateion between oral health and HCV infection[119], including a known association of HCV infection with oral lichen planus[120]. The influence of these features (oral health and smoking) on prediction of HCV infection are an area for further research.

Strengths and Limitations

Strengths of this study include the use of a current, nationally representative dataset that includes uninsured people and testing for HCV RNA of all participants, regardless of risk factor profile. Additionally, the use of ensemble methods for prediction of HCV infection achieved optimal performance in the training sample by combining multiple approaches (parametric and non-parametric) and using cross-validation to prevent over-fitting.

This study also had several limitations that should be considered when evaluating its findings.

Known limitations of NHANES data include the lack of inclusion of homeless persons, institutionalized populations (prisons, nursing homes, rehab facilities), and active military. Some of these unincluded groups are known to be at a higher risk of HCV than those in the general population[121, 122], and their omission may have had an impact on the list of characteristics identified by the prediction algorithm. Additional datasets should be identified to study the characteristics most predictive of HCV infection in these groups. Another limitation of the data involves the methods for capturing prescriptions. Studies have identified a trajectory from prescription opioid use to non-prescription opioid use through injection (which is a risk factor for HCV)[123]. NHANES reports on only those prescriptions that were taken in the 30 days prior to the household interview portion of the survey. It is likely that the 30-day window isn't sufficient to identify prescriptions that may have led to opioid addiction, injection drug use, and potential HCV infection.

We also excluded people under 20 years of age. Given the rising toll of the opioid epidemic in younger populations, there is evidence to suggest that searching for HCV infection in people under 20 may be important and fruitful[124].

As a cross-sectional study, there are inherent limitations to NHANES data, including: (1) difficulty establishing temporal ordering, (2) social desirability bias, and (3) non-response bias. NHANES data are collected in two encounters that are close in time (an interview and a physical

exam); therefore, it would be difficult to establish causal relationships between the various characteristics measured. As a result, this study focused on characteristics associated with HCV infection and did not attempt to draw conclusions about causality between these features and HCV infection. Additionally, NHANES tries to gather information about health-related conditions that may be sensitive or personal in nature. While an effort is made to create an environment that will result in honest responses, it is likely that some people were uncomfortable or unwilling to acknowledge or report behaviors that could be associated with an increased risk of HCV infection. Also important to consider is the impact of non-response bias. The NHANES survey involves time and effort, including agreeing to some modestly invasive testing (blood tests, physical exams). Participants are selected through a complex, multi-stage probability sampling design. In some cases, those selected do not complete both the interview and the exam or choose not to participate in certain portions of the exam. Differences between those who do and do not participate are unlikely to be completely random and could distort results.

Use of machine learning methods to predict a rare event presents its own set of challenges. The design of many machine learning algorithms includes an assumption of class balance in the outcome of interest. In the 2013 – 2018 NHANES sample, the prevalence of HCV infection was 1.1%. This degree of imbalance in individuals with and without the outcome could lead to poor performance in predicting the minority class because there are fewer examples from which the algorithm can learn and detect patterns. We attempted to address this problem by building our super learner algorithm with an unweighted case-control design that increased the concentration of HCV-infected individuals in the population sample, and then assessing performance with measures that emphasized positive case identification (i.e., using AUC-PR as opposed to AUROC, which rewards identification of true negatives). Finally, we report AUC-PR in both the case-control sample, and the full dataset with weights. In both cases, these measures improve precision over that which would result from universal testing at all levels of recall (sensitivity).

3.6: Conclusion

We developed a machine learning algorithm with a high level of precision to identify people with current HCV infection. Machine learning models excel at identifying complex relationships in high-dimensional data such as NHANES. A key contribution of this research was the identification of those features most influential for prediction of HCV infection. Expected findings included those related to measures of liver health (AST, ALT, HBV infection) and membership in the 1945 – 1965 birth cohort (age, triglycerides, sagittal abdominal diameter). Additional features that may merit further exploration included measures of primary and secondary cigarette smoke exposure and oral health.

Next steps for this research involve using the method and data described here to build a super learner with a more targeted suite of characteristics (possibly those available from a standard visit to a primary care provider) to create a portable tool for use in healthcare settings to help prioritize candidates for HCV screening. We believe the use of robust methods such as Super Learner, coupled with data collected from routine health care visits, may lead to novel pathways for diagnosing HCV infection and connecting individuals with treatment.

3.7 Figures and Tables

Figure 3.1 Flowchart of survey participant inclusion and exclusion criteria

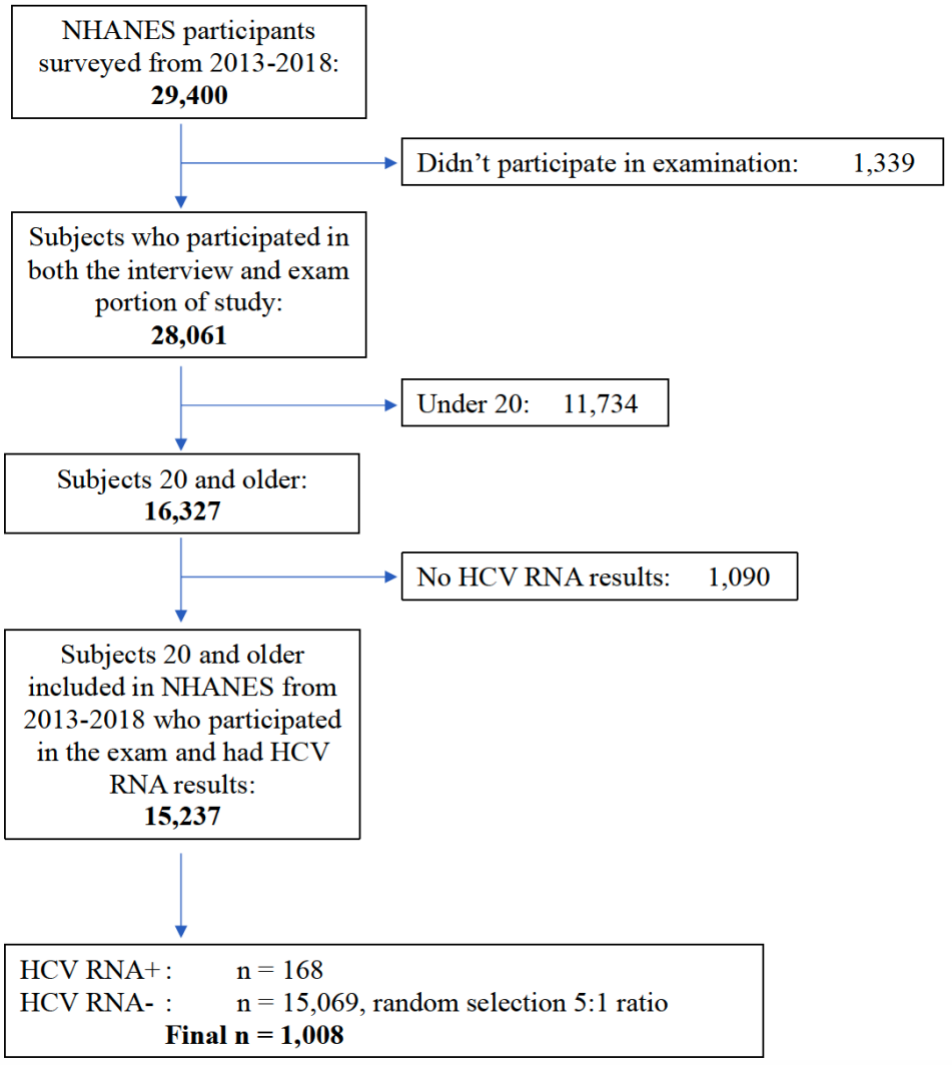


Figure 3.2: Feature exclusion and engineering

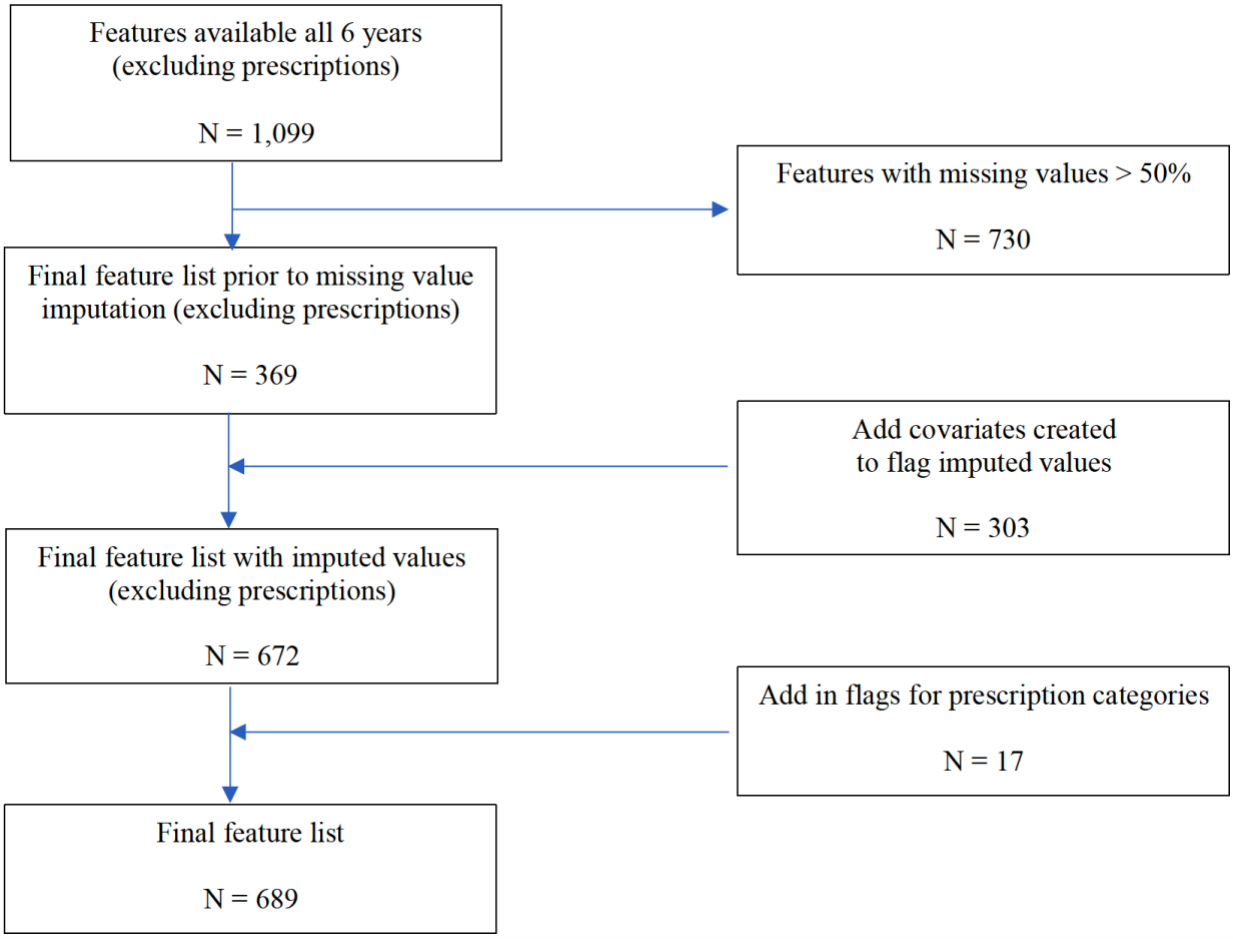
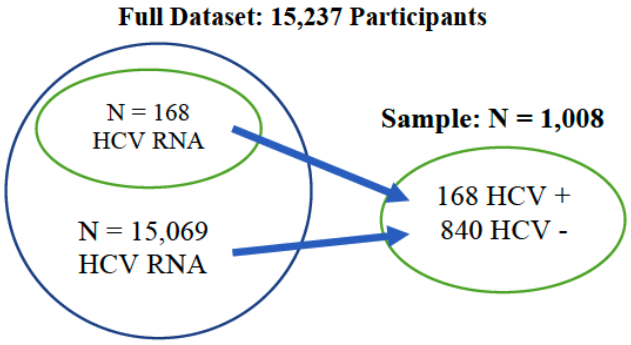


Figure 3.3: Model building

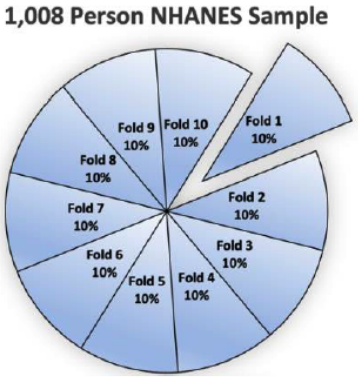
Step 1: Case Control Sampling

- Select all people who test positive for HCV RNA (N = 168)
- Randomly select five HCV-uninfected for each HCV-infected individual (5:1 Ratio; N = 840)



Step 2: Create Folds for Cross Validation

- Divide 1,008 rows into 10 folds
- Stratify folds by HCV infection status for balance



Step 3: Build Matrix of Out-Of-Sample Predictions

- Build a model for each learner on nine folds, predict on the 10th
- Repeat this process leaving out a different fold each time
- Build a matrix with
  - rows representing each learner
  - columns representing predictions

		1,008 People in NHANES Sample			
		Fold1 (Build Model: Folds 2-10)	**Fold2** (Build Model: Folds 1, 3-10)	**Fold3** (Build Model: Folds 1-2, 4-10)	**...** etc...
Individual Learners	XGBoost	Predict:XGBoost	Predict:XGBoost	Predict:XGBoost	...
	Ridge	Predict:Ridge	Predict:Ridge	Predict:Ridge	...
	LASSO	Predict:LASSO	Predict:LASSO	Predict:LASSO	...
	GLM	Predict:GLM	Predict:GLM	Predict:GLM	...
	etc...	...	...	...	...

Step 4: Feed Predictions to Super Learner

- Super Learner uses predictions in the matrix to create a weighted combination that performs at least as well as each individual learner

Figure 3.4: Precision recall curves for NHANES sample dataset and full dataset

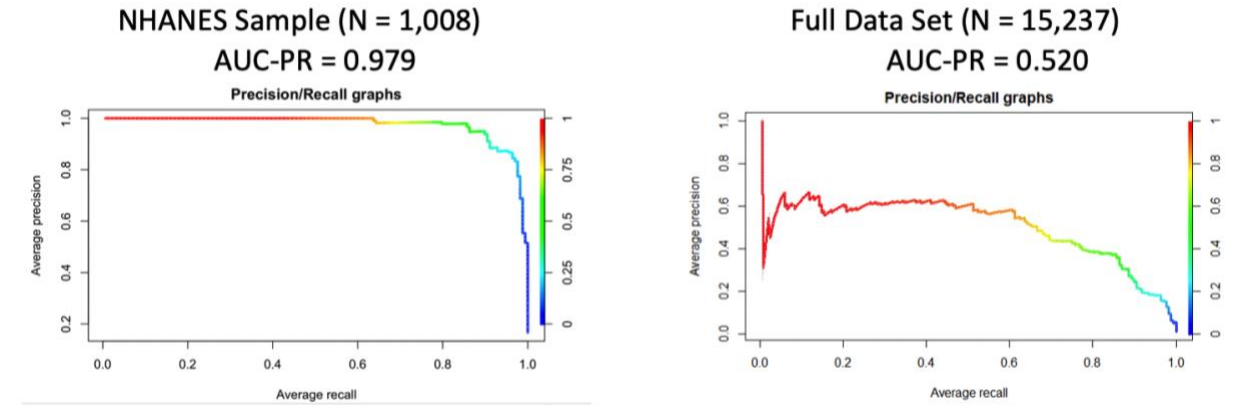
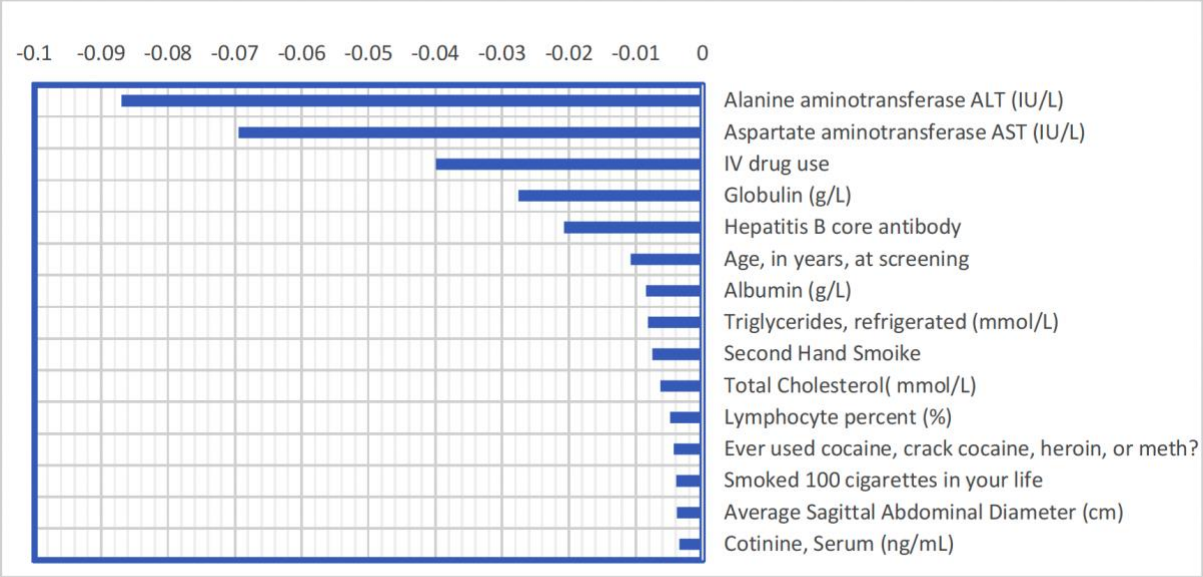
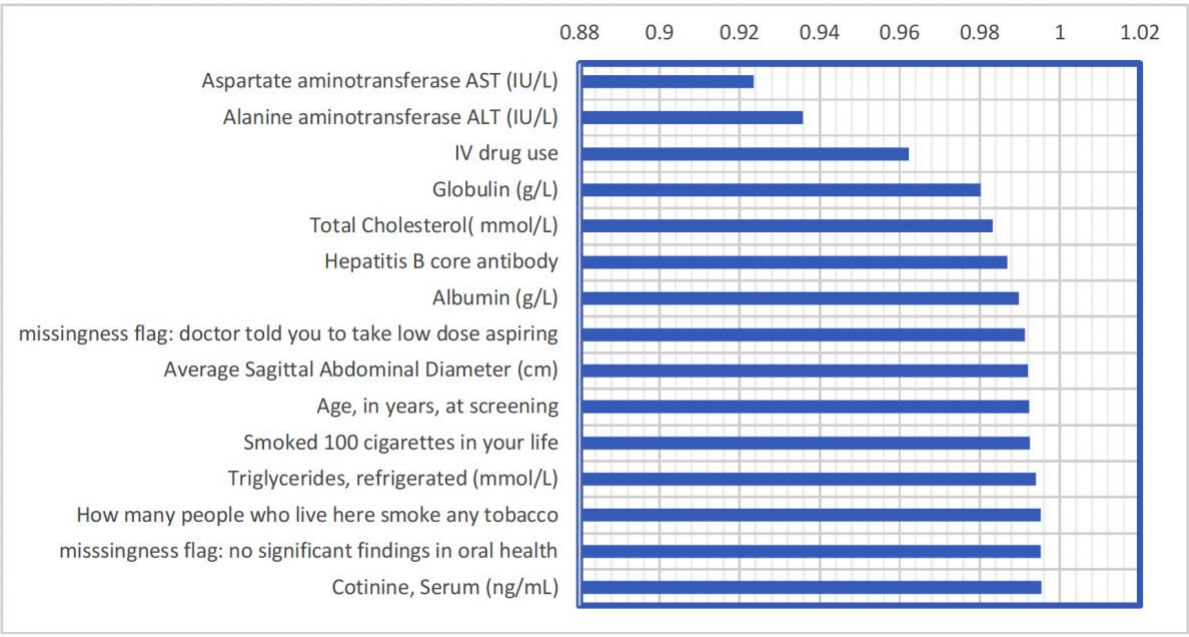


Figure 3.5: Top 15 influential features determined by the difference in area-under-the-precision-recall-curves without feature compared to including feature



* Note: a larger negative value is associated with greater influence on the AUC-PR

Figure 3.6: Top 15 influential features determined by the ratio of area-under-the-precision-recall-curve values without the feature compared to including the feature



* Note: a smaller fractional value is associated with greater influence on the AUC-PR

Table 3.1: Descriptive characteristics of NHANES sample (N= 1,008), stratified by HCV infection status

	HCV-Uninfected N (%)	HCV-Infected N (%)	p-value
N	840 (83.3)	168 (16.7)	
Age (years): Mean (SD)	50.1 (17.9)	56.7 (10.8)	< 0.001
Sex (% Male)	406 (48.3)	119 (70.8)	< 0.001
Race/Ethnicity			< 0.001
Non-Hispanic White	328 (39.0)	59 (35.1)	
Hispanic	213 (25.4)	32 (19.0)	
Non-Hispanic Black	174 (20.7)	64 (38.1)	
Non-Hispanic Asian	96 (11.4)	5 (3.0)	
Other including Multi-Racial	29 (3.5)	8 (4.8)	
College Educated	474 (56.4)	55 (32.7)	< 0.001
Married	491 (58.5)	65 (38.7)	< 0.001
Living at or below poverty level	165 (21.6)	76 (49.4)	< 0.001
HIV Positive	3 (0.5)	2 (2.2)	0.305
Injection Drug Use	14 (2.3)	66 (46.5)	< 0.001

Table 3.2: Feature pre-processing

Category	Starting Covariate Count (including restricted)	Publicly Released Covariates Available all Six Years	Notes
Demographic Features	52	16 (+4 admin)	Removed administrative and/or redundant
Examination	388	131	Aggregated dental variables
Questionnaire	1407 + (907 meds)	822	Aggregation, Convert to binary
Diet	910	-	Diet variables omitted
Laboratory	470	126	Removed redundant values, comment codes
Total	**4,134**	**1,099**	Total covariates included

Table 3.3: Area under the precision recall curve for individual learners and Super Learner, NHANES sample (N = 1,008)

	Original Fitted Super Learner				Cross Validated Super Learner			
Library of Learners	**coefficients**	**AUCPR**	**SE**	**Fold SD**	**coefficients**	**AUCPR**	**SE**	**Fold SD**
GLM	0.0909	0.8724	0.0145	0.0458	0.0909	0.8764	0.0159	0.0503
BayesGLM	0.0909	0.8735	0.0139	0.0441	0.0909	0.8778	0.0161	0.0509
Ranger	0.0909	0.8994	0.0152	0.0479	0.0909	0.8973	0.0148	0.0467
GLMNet, Alpha = 0.0 (ridge)	0.0909	0.8833	0.0195	0.0618	0.0909	0.8723	0.0257	0.0813
GLMNet, Alpha = 0.2	0.0909	0.8928	0.0170	0.0537	0.0909	0.8856	0.0199	0.0628
GLMNet, Alpha = 0.4	0.0909	0.8908	0.0170	0.0536	0.0909	0.8888	0.0176	0.0555
GLMNet, Alpha = 0.6	0.0909	0.8905	0.0165	0.0520	0.0909	0.8891	0.0169	0.0535
GLMNet, Alpha = 0.8	0.0909	0.8879	0.0165	0.0520	0.0909	0.8875	0.0170	0.0538
GLMNet, Alpha = 1.0 (LASSO)	0.0909	0.8872	0.0163	0.0514	0.0909	0.8869	0.0171	0.0540
XGBoost	0.0909	0.8919	0.0126	0.0399	0.0909	0.8978	0.0120	0.0380
Dbarts	0.0909	0.8959	0.0168	0.0533	0.0909	0.8928	0.0185	0.0586
Superlearner	***NA***	**0.9012**	**0.0139**	**0.0439**	***NA***	**0.9051**	**0.0140**	**0.0442**

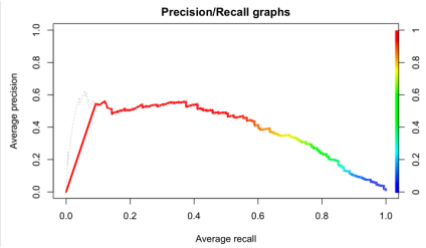
Table 3.4: Performance Metrics on the NHANES Sample and on the Full Data Set

Metric	NHANES Sample (N=1,008)	Full Data Set (N = 15,237)
Baseline (HCV Prevalence)	0.167	0.011
Prediction threshold to maximize F1*	0.382	0.854
Max F1 score	0.924	0.582
Precision	0.944	0.554
Recall	0.905	0.613
Accuracy	0.975	0.990
Specificity	0.989	0.994
AUC	0.995	0.989
AUC-PR	0.979	0.520

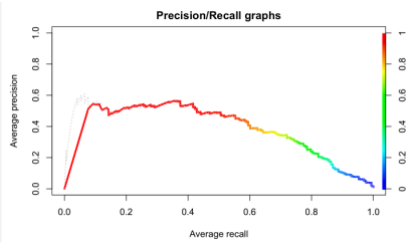
3.8: Supplemental Material

Supplemental Figure S3.1: Area under the precision recall curve graphs for individual learners on full data set (N = 15,237)

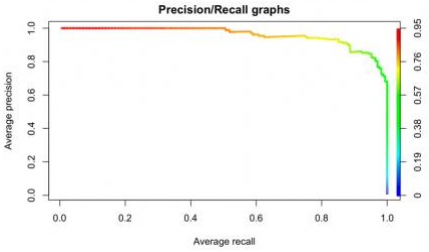
GLM (AUC-PR = 0.401)



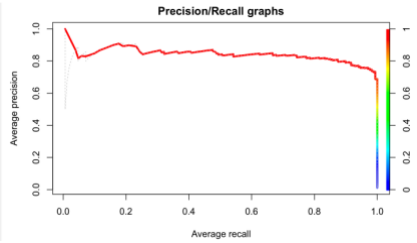
BayesGLM (AUC-PR = 0.401)



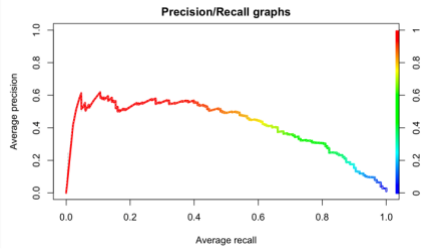
Ranger (AUC-PR = 0.961)



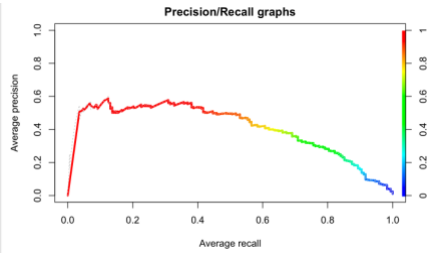
XGBoost (AUC-PR = 0.840)



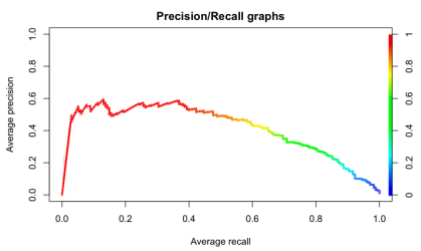
Ridge, alpha = 0 (AUC-PR = 0.426)



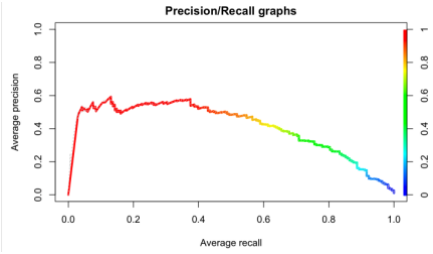
Lasso, alpha = 1 (AUC-PR = 0.414)



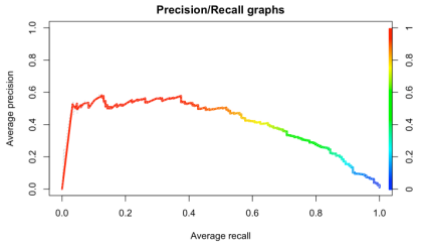
Elastic Net, alpha = 0.2 (AUC-PR = 0.421)



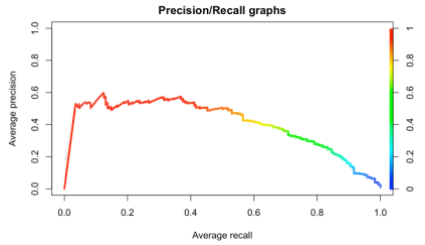
Elastic Net, alpha = 0.4 (AUC-PR = 0.417)



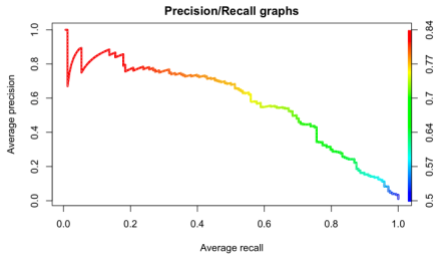
Elastic Net, alpha = 0.6 (AUC-PR = 0.416)



Elastic Net, alpha = 0.8 (AUC-PR = 0.415)



dbarts (AUC-PR = 0.584)



Chapter 4: Predicting hepatitis C virus infection: use of Super Learner with data from the National Health and Nutrition Examination Survey to find individuals with undiagnosed HCV infection

4.1: Abstract

Hepatitis C virus (HCV) infection is a global health concern and the source of significant morbidity and mortality. Of the estimated 2.4 million individuals in the United States (US) infected with HCV, approximately half are unaware of their infection. Proactive case-finding is necessary to achieve the WHO sustainable development goal of viral hepatitis elimination by 2030. To support this effort in the US, we used national survey data and machine learning methods to develop an HCV prediction algorithm to identify individuals with a high probability of HCV infection who should be prioritized for HCV screening.

Super Learner (SL) is a loss-based ensemble learning method that uses cross-validation to estimate the performance of multiple machine learning models and creates an optimal weighted average of those models using test data performance. Our prediction algorithm was built using National Health and Nutrition Examination Survey (NHANES) data collected from 2013 – 2018 and a diverse library of parametric and non-parametric classifiers and screening algorithms to optimize performance and model run time. NHANES was selected because it employs a multistage probability sampling design to select participants and includes universal testing for HCV RNA. Class imbalance is present in the dataset due to the low prevalence of HCV infection in the US population (1-2%). We addressed this issue by using synthetic minority over-sampling technique (SMOTE) with a random sample of controls from the full NHANES dataset and a meta-learner that maximizes area under the precision-recall curve (AUC-PR). Experts were consulted to identify an appropriate subset of NHANES features (independent variables such as laboratory measures that are commonly available in electronic medical records) to build the algorithm. In addition, a probability threshold was selected to maximize the F1 score (i.e. the harmonic mean of precision and recall).

Using the fitted SL, the maximum F1 score 0.531 was achieved with a threshold value of 0.859, meaning any individual whose predicted probability of HCV infection was 85.9% or higher was classified as a positive case. At this threshold, the prediction algorithm achieved 21.8% precision and 86.3% recall, and an AUC-PR of 47.7%. These results can be interpreted to mean that for every 100 individuals flagged as HCV positive by this algorithm, approximately 21 individuals would be true positives, and more than 85% of individuals with HCV infection would be identified. AUC-PR is a score representing success at predicting positive cases, where no-skill would be the underlying prevalence of HCV in the data (in this case, 1.1%). The AUC-PR of 47.7% represents a 43-fold improvement over baseline. When the fitted SL was applied to the full NHANES dataset from 2013 – 2018, including survey weights, precision was maintained close to 55% at recall levels up to 60%.

An HCV infection prediction algorithm developed with machine learning methods and commonly available laboratory measures can support early identification of HCV infection by

flagging individuals with a high probability of infection to enable prioritized HCV screening. Earlier diagnoses will enable earlier clinical interventions and improved HCV outcomes.

4.2: Introduction

Hepatitis C virus (HCV) infection, a bloodborne infection that can cause inflammation and progressive scarring of the liver, is a significant public health concern worldwide. While 30% (15-45%) of people infected with HCV will spontaneously clear HCV in the first six months[10], for most the infection becomes chronic and is associated with long term hepatic and extrahepatic complications, including cirrhosis, liver failure, and hepatocellular carcinoma[125-128]. In 2015, approximately 71 million people worldwide and 2.4 million in the US were chronically infected with HCV, producing an estimated 400,000 deaths globally and 15,000 in the US due to HCV-related sequelae. These estimates are considered conservative due to under-reporting and under-diagnosis of HCV infection[57, 129, 130].

Historically, treatment of HCV infection involved administering weekly injections of interferon for up to 48 weeks. Interferon treatment provided only limited success in suppressing HCV infection and did not result in elimination of the virus, and at the same time was associated with important side effects in many patients[131, 132]. The 2013 approval of highly efficacious direct acting antiviral (DAA) treatment brought about a paradigm shift in treatment of HCV infection, transforming the landscape from one where HCV was managed to one where it could be effectively eliminated in most patients[15, 16]. However, HCV infection is generally asymptomatic until significant liver damage has occurred, presenting a major obstacle to diagnosis and treatment. Up to 90% of people globally and 50% of people in the US who are infected with HCV are unaware of the infection[18, 19, 102].

In 2015, the United Nations (UN) established 17 Sustainable Development Goals (SDGs). Key among these were the prioritization of health and well-being, with a specific target to eliminate viral hepatitis by 2030. In 2016, the World Health Assembly (WHA), the decision-making arm of the World Health Organization (WHO) comprised of health ministers from 194 member-states, established goals that aligned with the SDGs by approving the Global Health Sector Strategy on Viral Hepatitis. The WHO hepatitis strategy aims to eliminate viral hepatitis by 2030, with specific goals to diagnose 90% of those with HCV infection, reduce incidence by 80% and reduce mortality by 65% [56].

Though the United States (US) has adopted one-time universal HCV screening for all adults 18-79 years of age[98], the diagnosis and treatment milestones needed to achieve the WHO's 2030 global elimination goals are not being met. A recent study of progress toward HCV elimination in 45 countries found that, at the current rate of diagnosis and treatment, only 24% of countries were on target to achieve HCV elimination by 2030, and more than 50% of countries would not achieve elimination by 2050, including the US[64]. Further, these elimination estimates are likely optimistic, given the disruption to healthcare systems and outreach programs caused by the global Covid-19 pandemic.

In light of slower-than-hoped-for progress in identifying and treating people with HCV infection in the US, strategies to proactively search for individuals infected with HCV are needed more

than ever. One approach to this problem involves the use of machine learning methods together with available healthcare datasets to predict HCV infection status and prioritize individuals for HCV screening.

Machine learning methods are particularly well-suited to identifying complex, non-linear patterns in high dimensional data, such as the data captured in electronic medical records (EMR) and claims data.[104, 105] Studies comparing the prognostic abilities of machine learning methods to those of traditional models have found better predictive performance by the machine learning models, particularly when a greater number of predictors were used[133, 134].

We consulted with primary care physicians and experts on EMR data analytics to identify a set of laboratory values and demographic characteristics ("features") that are reliably captured and documented during healthcare encounters. Features were selected based on the following criteria:

- likely to be available with a high degree of completeness for most patients after an annual physical exam or encounter with a primary care provider
- available in structured data (i.e. not from free-text fields, which pose analytic challenges).

From these identified features, we used ensemble machine learning methods to build an HCV infection prediction algorithm that could be applied to a US EMR database to identify people at high risk for HCV infection who should be prioritized for HCV screening.

Building a prediction algorithm using supervised learning requires the use of labeled data – that is, data in which the HCV infection status of the included individuals is known. With this in mind, we used information from the National Health and Nutrition Examination Survey (NHANES). NHANES is an annual survey that has been conducted in the US since the late 1960s. It includes demographic characteristics and laboratory measures from a multi-stage randomly selected sample of participants in the un-institutionalized and housed US population. All participants in NHANES are tested for HCV RNA and a full suite of additional laboratory measures, regardless of observable risk factors.

The objective of this study was to use ensemble machine learning methods to develop an HCV infection prediction algorithm that can be applied to electronic health records to identify individuals at high risk for HCV infection and prioritize them for screening. Using the three most recent cycles of NHANES (2013 – 2018), we identified the demographic and laboratory information that would likely be available and complete in an EMR dataset. With these data and a diverse suite of prediction algorithms, we created a model with 10-fold cross validation to predict HCV infection. This model was assessed by calculating the area under the precision recall curve (AUC-PR), a metric that is preferred when the outcome of interest is rare (< 10% of the total population) because it focuses on the correct identification of positive cases.

4.3: Methods

Data Source

The National Center for Health Statistics (NCHS) is a branch of the US Centers for Disease Control and Prevention (CDC), whose mission is to collect data and disseminate statistical findings to inform public health decision-making. One of the major programs of the NCHS is the NHANES, an annual survey conducted with the intention of taking a snapshot of the nation's

health. NHANES was established in the 1960s, and has collected data continuously since 1999, gathering health and nutrition measurements from individuals randomly selected using a complex multi-stage sampling design in 15 US counties each year.

An initial interview is conducted at the participant's home to gather demographic, diet, and socioeconomic information, as well as a medical history. The second part of the survey includes a physical examination at a mobile examination center (MEC), where blood, urine, and other biological specimens are collected and processed. All laboratory values and demographic information collected by NHANES were considered for inclusion in this analysis.

NCHS Research Ethics Review Board (ERB) approval to conduct NHANES was obtained for data collected from 2013-2016 under a continuation of protocol #2011-17. Approval for data collection in 2017-2018 was approved under protocol #2018-01. All NHANES data are de-identified and publicly available on the NCHS and USDA websites. As such, this study was exempt from institutional review board/ethics review board oversight[135].

Study Cohort

The overarching goal of this study was to develop a tool to identify individuals who have a high probability of infection with HCV so they can be prioritized for screening. Given the changing demographic features of the HCV-infected population, driven in part by the current opioid epidemic, it was determined that the most effective prediction algorithm would be built using contemporary demographic and medical records. To this end, cohort inclusion was restricted to participants in the three most recent NHANES cycles: 2013-2014, 2015-2016, and 2017-2018. In addition, participants were retained in the study cohort only if they met all the following inclusion criteria:

- completed the medical examination portion of the survey
- 20 years of age or older
- HCV ribonucleic acid (RNA) test result available

Participation in the medical examination was required for inclusion because laboratory specimens were collected during this portion of the survey. Many laboratory measurements, however, were collected only for NHANES participants 20 years of age and older, so participants younger than 20 years were excluded from the current study. Last, because this study employed supervised learning methods to identify patterns of characteristics associated with HCV infection status, labeled training data (i.e. data where HCV RNA status is known) were necessary. As a result, only NHANES participants with non-missing HCV RNA measurements were included.

Feature Selection

Patient information available in an EMR has both structured and unstructured elements. Structured data are stored and organized in a consistent and predictable format that minimizes the amount of pre-processing necessary for sorting, combining, and analyzing. Unstructured data can appear in a variety of formats and generally require significant cleaning and interpretation to achieve an analyzable structure. In this study, we chose to build the best possible HCV infection prediction algorithm from data that would be available in a structured format in an EMR data set to minimize researcher assumptions and simplify study replication in NHANES and other data

sets. To this end, we have included only quantitative laboratory measures and minimal demographic features in our analytic data set.

The dependent (predicted) variable in this analysis is a binary flag indicating HCV RNA status for each participant. NHANES uses a nucleic acid amplification test (the COBAS AmpliPrep/COBAS TaqMan HCV Test) to assess HCV RNA from a blood sample[136]. NHANES participants who did not have a result for this test were excluded. The independent variables (aka predictors or features) were selected from the full list of demographic and laboratory data present across all three NHANES cycles from 2013 – 2018. Any predictor in NHANES with missing values for ≥ 50% of participants was excluded. We consulted with subject matter experts and EMR data analysts to identify the predictors from those remaining that would be captured as quantitative (structured) measures from a physical or primary care appointment with a high degree of completeness and reliability in EMR data.

Sampling Method

Many machine learning algorithms used for binary classification have been designed with the assumption that the data on which the classifier is built contain relatively equal numbers of cases and non-cases. Given that only approximately 1% of NHANES participants test positive for HCV RNA, this assumption does not hold and could result in a model that does a poor job recognizing the minority class (HCV RNA positive). No universal solution has been developed to address prediction with class imbalance. However, a variety of different approaches have been devised to improve prediction performance, with the best solution being dependent on the specifics of the data used to train the model.

We considered several sampling approaches and selected the one which consistently yielded the highest cross-validated AUC-PR. The approaches investigated included:

1) Use of the full dataset: Data from all available participants were used to build the prediction algorithm. The prevalence of HCV positive individuals in this dataset was 1.1%.
2) Case-Control Sampling: Data from all participants who tested positive for HCV RNA were included to build the prediction algorithm, along with a randomly selected subset of individuals who tested negative for HCV RNA (5 “non-cases” for every “case”). As a result, the prevalence of individuals with HCV infection was now 16.7%.
3) Synthetic Minority Over-Sampling Technique (SMOTE): In this sampling method, the minority class (those with HCV infection) was augmented to establish equal prevalence (50%) in the dataset, enabling improved performance from the machine learning algorithms. Oversampling of the minority class was achieved through the creation of synthetic data generated by randomly choosing points on the lines that connected rare observations to a user-specified number of nearest neighbors in feature space, rather than simply sampling rare cases with replacement[137].
4) Random Over-Sampling Examples (ROSE): To create balance in the classes of the dependent variable, the minority class was oversampled through the creation of synthetic data generated based on the conditional density of the underlying data and the majority class was under-sampled through a process of random selection to create an appropriate-sized subset[138].

Statistical Analysis

We used the ensemble learning algorithm Super Learner[139] (SL) with 10-fold cross-validation to build our predictor. Ensemble learning considers a user-specified library of candidate prediction algorithms with a process called stacking (individual algorithms are trained simultaneously, allowing for comparing and combining) and creates a weighted combination from the output of the individual learners that minimizes the cross-validated empirical risk associated with the user-specified loss function. We employed screening algorithms with each learner to optimize performance and model run time. Screening algorithms filter out features that aren't significantly contributing to successful prediction prior to the building of the final model.

A diverse library of parametric and non-parametric algorithms was selected to maximize the predictive power of the final model. Parametric learners are faster, require less information, and perform well with simple prediction problems, whereas the non-parametric learners have greater flexibility to identify complex relationships and don't fall prey to misspecification:

Parametric Learners

- Generalized linear model (*glm*): GLM fits a standard generalized linear model
- *Bayesglm*: Bayesglm is an alternative to GLM that uses student-t prior distributions for the coefficients to produce more stable estimates. In a model with many features, especially those with low variance, the use of priors can significantly reduce efficiency. Bayesglm uses a modified expectation-maximization algorithm to fit the model[108, 109].
- *Glmnet*: Glmnet estimates a regularized generalized linear model. Penalty options include the range from l1 (Least Absolute Shrinkage and Selection Operator; *LASSO*), which can shrink the slope of unhelpful coefficients to 0, effectively removing them from the model, to l2 (ridge), which keeps all features but shrinks their slopes to close to 0, and mixtures of LASSO and ridge (elastic net). The mixture is established by a specified alpha (α) that ranges from 0 (ridge) to 1 (LASSO)[110]. We used glmnet learners with alphas ranging from α= [0, 0.2, 0.4, 0.6, 0.8, 1].

Non-Parametric Learners:

- Extreme gradient boosting (*xgboost*)[111]: Xgboost uses gradient boosted decision trees and is optimized for speed and model performance. Boosting builds models in a sequential manner and uses a loss function and weights to focus on the most challenging cases (sequentially higher weights are given to misclassifications for subsequent iterations). The number of fitting iterations was set to 500, and the number of early stopping rounds was set to 50. Early stopping, a way to avoid over-fitting, is used when the loss on the validation set starts to increase.
- Discrete Bayesian additive regression tree sample (*dbarts*)[112]: Dbarts is a Bayesian tree ensemble method that uses individual trees as base learners. Each tree is constrained by a prior to be a weak learner. This learner is flexible and requires minimal assumptions.

  Ranger: Random forest (an ensemble method using decision trees and bagging) is optimized for high dimensional data[113]. Ranger builds on random forest by allowing the user to choose a mode for calculating variable importance, the contribution a specific feature makes to prediction. Importance criterion "impurity" was used

(impurity measurement uses the Gini index[114] for classification that is the probability that a randomly selected feature is classified incorrectly).

Meta-learning refers to the process of learning from the output of the individual learners and combining that output with a specified goal. We defined a metalearner with nonlinear optimization via augmented lagrange and selected an evaluation function that maximized area AUC-PR. This type of metalearner has been shown to outperform metalearners specified with other loss functions, as well as outperforming individual machine learning algorithms[140].

The initial fit for the Super Learner was performed using 10-fold cross validation. The data were divided into ten folds, each with 10% of the NHANES participants – stratified so each fold contained equal numbers of individuals with HCV infection, with no duplication. Each individual learner was trained on nine folds, then validated on the 10^{th} fold, which was unseen during model training. This process was repeated, holding out a different fold each time until every individual in the dataset had an HCV prediction for all individual learners based on an algorithm that was not trained with their data. The final fitted Super Learner was the product of a weighted combination of the HCV infection predictions from the individual learners.

Performance of the fitted Super Learner was assessed on unseen data by again using 10-fold cross validation (recreating the Super Learner on 90% of the data, validating on a 10% holdout sample, and repeating the process ten times, leaving out a different 10% sample each time). The final fitted Super Learner was deployed to predict HCV infection status on the full NHANES dataset, using NHANES-supplied weights to accurately represent the US population surveyed during the six-year period (2013 – 2018).

Performance Metrics

Performance characteristics for individual learners and for the Super Learner were assessed by AUC-PR and the F1-score.

The precision-recall curve is made up of precision on the y-axis and recall on the x-axis and shows their relationship at all possible prediction thresholds. Precision, also known as positive predictive value (PPV) is defined as the proportion of successfully identified positive cases (TP) among the total number of successfully predicted positives (TP and false positives; FP). Recall, also known as sensitivity, is defined as the proportion of TP among the total population of positive cases (TP and false negatives; FN).

$$\text{Precision: } \frac{TP}{(TP+FP)} \qquad \text{Recall: } \frac{TP}{(TP+FN)}$$

The AUC-PR provides a way to quantify successful prediction of positive cases. This metric is desirable when the outcome of interest is rare because it doesn't reward successful prediction of the true negatives that make up the majority of the data.

The geometric interpretation of the precision recall curve is the expected precision when uniformly varying the recall. Limitations of precision-recall curves include

- lack of universal baseline

- uninterpretable region on the lower right side of the graph
- lack of calibration [115]

Precision and recall can be assessed through the F-score (F combines precision and recall into one metric and is more useful than accuracy when you have class imbalance), where, when β = 1, F is the harmonic mean of precision and recall[116]. The value of β determines the tradeoff between precision and recall. When $\beta > 1$, recall is given greater weight; when $\beta < 1$, precision receives more weight.

$$Accuracy = \frac{True}{True + False} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F_\beta = (1 + \beta^2) * \frac{Precision * Recall}{(\beta^2 * Precision) + Recall}$$

The baseline in a precision recall curve is equivalent to the true prevalence of the outcome at all values of recall (forming a horizontal line). At maximum sensitivity, all samples are predicted to be cases.

4.4: Results

Descriptive Analysis

Participants in the three most recent NHANES cycles from 2013 – 2018 (n = 29,400) were evaluated for inclusion in this study. Among them, 1,339 did not complete the medical examination portion of the survey. Of the remaining 28,061participants, 11,734 were under 20 years of age, and 1,090 did not have a recorded measurement for HCV RNA. After excluding those groups,15,237 individuals remained in the cohort (**Figure 4.1**), with an HCV RNA positivity prevalence of 1.1% (n = 168).

Demographic and key laboratory values associated with HCV infection status are provided in **Table 4.1**. In the unweighted NHANES cohort, individuals who tested positive for HCV RNA were older (mean age 56.7 years; standard deviation (SD) 10.8), and more likely to be male, born in the United States (US), living below the poverty level, and coinfected with hepatitis B virus (HBV) and human immunodeficiency virus (HIV), whereas people who tested negative were younger (mean age 50 years; SD 17.7) and more likely to be college educated and married.

Of the 126 laboratory measures and 22 demographic characteristics available in NHANES from 2013 – 2018, 82 were excluded due to poor availability in EMR data. Sixteen demographic features that would not be reliably captured in structured fields were also excluded. Among the 49 remaining features, none had missing values for ≥ 50% of NHANES participants, so no additional features were excluded; however, missing values were imputed in 40 of the 49 features. The final dataset used by Super Learner included 49 features and 40 indicator variables with missingness flags for each variable that contained missing values requiring imputation (**Figure 4.2**).

Four sampling approaches were evaluated when building the HCV infection prediction algorithm. The SMOTE sampling method yielded the highest cross-validated AUC-PR. Results from all sampling methods are available in **Supplemental Table S4.1**. With SMOTE, laboratory results and demographic characteristics from the 168 individuals who tested positive for HCV RNA were used to generate synthetic data representing 15,042 HCV infected "cases" (**Figure 4.1**). These synthetic HCV positive cases balance out the 15,069 individuals who tested negative for HCV RNA, which gives the model equal opportunity to learn about patterns of HCV positivity.

Prediction Results:
HCV infection prediction results for the eleven individual learners and for the final Super Learner are presented in **Table 4.2**. Employing a dataset with 89 features (49 discrete predictors and 40 indicator variables for missing values in specific predictors) and using augmentation of the minority class (individuals with HCV infection) via SMOTE to produce 15,042 synthetic cases of HCV infection to accompany the 15,069 individuals without HCV RNA present, the individual learners and the Super Learner achieved almost perfect discrimination in identifying individuals with and without HCV infection. A diagram of the model building process is presented in **Figure 4.3**. Performance for each learner was evaluated based on its ability to maximize AUC-PR. Using this metric, scores ranged from 0.9984 (standard error (SE) 0.0004) for both glm and Bayesglm up to 1.0000 (SE 0.0000) for ranger.

Non-parametric learners (dbarts, xgboost, and ranger) outperformed the parametric learners in both the initial fit of the Super Learner (with 10-fold cross validation) and in the cross-validation of the Super Learner itself (again with 10-fold cross validation). The Super Learner also outperformed the parametric learners with an AUC-PR of 0.9995. The ensemble learner is a combination of equally weighted (0.0909) individual learners. Reasons for the identical contribution of each learner are unclear; however, it may be related to fact that individual learners performed similarly.

The final fitted Super Learner was deployed on the unweighted, SMOTE-augmented dataset (HCV = 15,042; non-HCV = 15,069) and then on the full, weighted NHANES dataset (HCV = 168, non-HCV = 15.069) to predict HCV infection. Performance metrics evaluating HCV infection prediction success are presented in **Table 4.3**.

In the unweighted, SMOTE-augmented dataset, the prevalence of HCV infection was approximately 50%. To identify the probability threshold that maximized the F1 score, values from 0.001 to 0.999 were tested. A maximum F1 values of 0.976 was achieved with a threshold of 0.513, meaning that any individual with the probability of HCV infection at or above 51.3% was flagged as a positive case. With this threshold, the prediction algorithm achieved 99.4% precision at 50.8% recall and an AUC-PR of 99.1%.

In the weighted full NHANES dataset, the prevalence of HCV infection was approximately 1.1%. Using the fitted Super Learner, the maximum F1 score of 0.531 was achieved with a threshold value of 0.859, meaning any individual whose predicted probability of HCV infection was 85.9% or higher was classified as a positive case. At this threshold, the prediction algorithm achieved 21.8% precision and 86.3% recall, and an AUC-PR of 47.7%, a 43-fold improvement

over baseline. These results can be interpreted to mean that for every 100 individuals flagged as HCV positive by this algorithm, approximately 21 individuals would be true positives, and more than 85% of individuals with HCV infection would be identified.

Precision-recall curves for the SMOTE-augmented Super Learner and for the full weighted NHANES Super Learner are presented in **Figure 4.4**. Using the dataset on which the Super Learner was trained, precision remained at 99% for almost all levels of recall. When the Super Learner was applied to the full NHANES dataset from 2013 – 2018, including survey weights, precision was maintained close to 55% at recall levels up to 60%. For example, at 20% recall, for every 100 patients flagged by the algorithm, approximately 60 would be true positives, and at 50% recall, for every 100 patients flagged by the algorithm, 50 would be true positives. As recall goes above 60%, the decline in precision becomes more pronounced.

4.5: Discussion

Achieving the WHO's goal to eliminate viral hepatitis by 2030 will require a multi-pronged approach, and HCV prediction algorithms can play a crucial role in efficiently identifying the target population for screening. Globally, WHO estimates that only one out of every 10 people infected with HCV are aware of their infection[56]. In this study, we demonstrated that machine learning methods can be employed to improve the identification of individuals with HCV infection, as compared to relying exclusively on one-time universal HCV screening. The use of rich data sources, such as national health survey data and medical records/claims data, coupled with sophisticated tools like Super Learner, can optimize screening resources by prioritizing individuals who have a high likelihood of being infected with HCV. Diagnosing and treating as many HCV-infected individuals as possible will both reduce sequelae of HCV and prevent transmission from infected individuals to others.

In this study, the Super Learner algorithm trained on SMOTE-augmented data achieved almost perfect discrimination in out-of-sample prediction. The analytic data set included every NHANES participant in the cohort who tested negative for HCV RNA and synthetic data representing a relatively equal number of HCV positive NHANES participants. Augmenting the data to include equal representation for people with and without HCV helps to avoid prediction bias toward the better-represented class[141]. Prediction performance from the algorithm built with the SMOTE-augmented data significantly outperformed the algorithms developed on the full NHANES dataset and on a case-control subset of the data, which included all 168 individuals with HCV infection and a random 5:1 sample of individuals without HCV infection (Supplemental Table 1). A second minority class oversampling technique, ROSE, was also tested and achieved perfect discrimination in out-of-sample prediction, but it performed no better than baseline (HCV positivity prevalence) when deployed on the full NHANES weighted dataset. Reasons for this poor performance are unclear.

The AUC-PR of the prediction algorithm described here (46.1%) demonstrates a 42-fold improvement over the AUC-PR associated with universal screening (equivalent to a population prevalence of HCV infection of 1.1%). There is an inverse relationship between precision and recall, meaning as recall is increased, precision will decrease. When graphed together, the slope of the relationship between these two measures is not necessarily constant. The precision of the

Super Learner created with the SMOTE dataset remained relatively constant for recall levels up to 60%, and then declines steeply as recall increases above that. Different thresholds can be considered in the prediction algorithm, depending on tolerance for false positive predictions weighed against identifying as many individuals with HCV infection as possible. At maximum recall, the algorithm worked no better than universal screening.

Previous studies have employed machine learning to support the identification of individuals with HCV infection. In one study, Doyle et al. demonstrated that ensemble learning methods can help diagnose individuals with HCV infection earlier [104]. Using longitudinal administrative claims data, their findings suggest that patients with an HCV diagnosis or evidence of treatment had claims for HCV-related symptoms up to two to three years prior to the presence of HCV diagnosis codes or HCV treatment dispensing codes. A recent study by Orooji et al. described the design of an HCV infection prediction algorithm using machine learning in a dataset with significant class imbalance. They reported that performance measures for their algorithm were improved when they used over- and under-sampling techniques to create balance[142]. Our results confirmed these findings by successfully leveraging the oversampling technique to create class balance and by using ensemble learning methods to improve HCV infection identification over universal screening. Our findings also advance efforts to identify the HCV-infected population by training the prediction algorithm on a large US population-based random sample of individuals with known HCV infection status.

Strengths and Limitations

This study combines six years of contemporary survey data, including the most recent results of 2017-2018 NHANES. To our knowledge, no published study of HCV prediction has utilized these data. NHANES reflects one of the most up-to-date and thorough snapshots of the health of non-institutionalized and housed US residents.

Another strength of our study is the focus on structured data for prediction. Data available in free text fields present a unique analytic challenge in terms of abstraction and aggregation. Misspellings, difficulty in interpreting handwriting in scanned documents, differences in naming and abbreviation conventions, and transcription errors create significant challenges for consistently capturing and interpreting medical information. Structured data, such as numeric results from laboratory tests, require less pre-processing and are easier to reliably aggregate and analyze than data extracted from free-text fields.

A major strength of the NHANES dataset is that HCV status is determined for all who participate in the examination portion of the survey. In claims and EMR data, the cohort of individuals who do not have an HCV diagnosis includes two types of people – people who have HCV infection but have not been diagnosed, and people who do not have HCV infection. Unfortunately, the comingling of these two populations adds significant noise to the development of a classification algorithm and could potentially perpetuate the underdiagnosis of HCV infection. In NHANES data, the non-HCV cohort contains only individuals who have tested negative for HCV RNA. As a result, a classification algorithm that learns from NHANES data will be able to identify patterns associated with individuals who are truly HCV RNA negative, rather than from the combination of true negatives and undiagnosed positives.

The cross-sectional study design of NHANES presents a limitation for the portability of the prediction algorithm. Deploying this HCV infection prediction algorithm in longitudinal data, such as EMR or claims data, will require decisions about how to handle laboratory findings that occur at different points in time. In addition, NHANES data include most laboratory values for every participant; however, EMR and claims data will likely have data sparsity which may limit the number of individuals who have sufficient data to receive an HCV infection classification.

Next Steps

The longterm goal of the work described herein is to develop a successful HCV infection prediction algorithm that can identify high priority screening candidates and contribute meaningfully to the WHO goal of HCV elimination by 2030. Our algorithm has been validated on de-identified, cross-sectional data where the HCV infection status of included individuals was already known. Next steps will include partnering with a health care system to pilot the algorithm on EMR data to identify high risk candidates for HCV infection. Ideally those patients flagged by the algorithm will be screened for HCV, to determine the real-world success of this prediction algorithm.

4.6: Conclusion

As demonstrated in our analysis, an HCV infection prediction algorithm developed with machine learning methods and commonly available laboratory measures can support early identification of HCV-infected individuals. Earlier diagnoses will enable earlier interventions and improved HCV outcomes. As healthcare systems adopt expanded HCV screening recommendations, this type of prediction algorithm can be used to focus resources on the population most likely to benefit.

4.7: Figures and Tables

Figure 4.1: Flowchart of survey participant inclusion and exclusion criteria

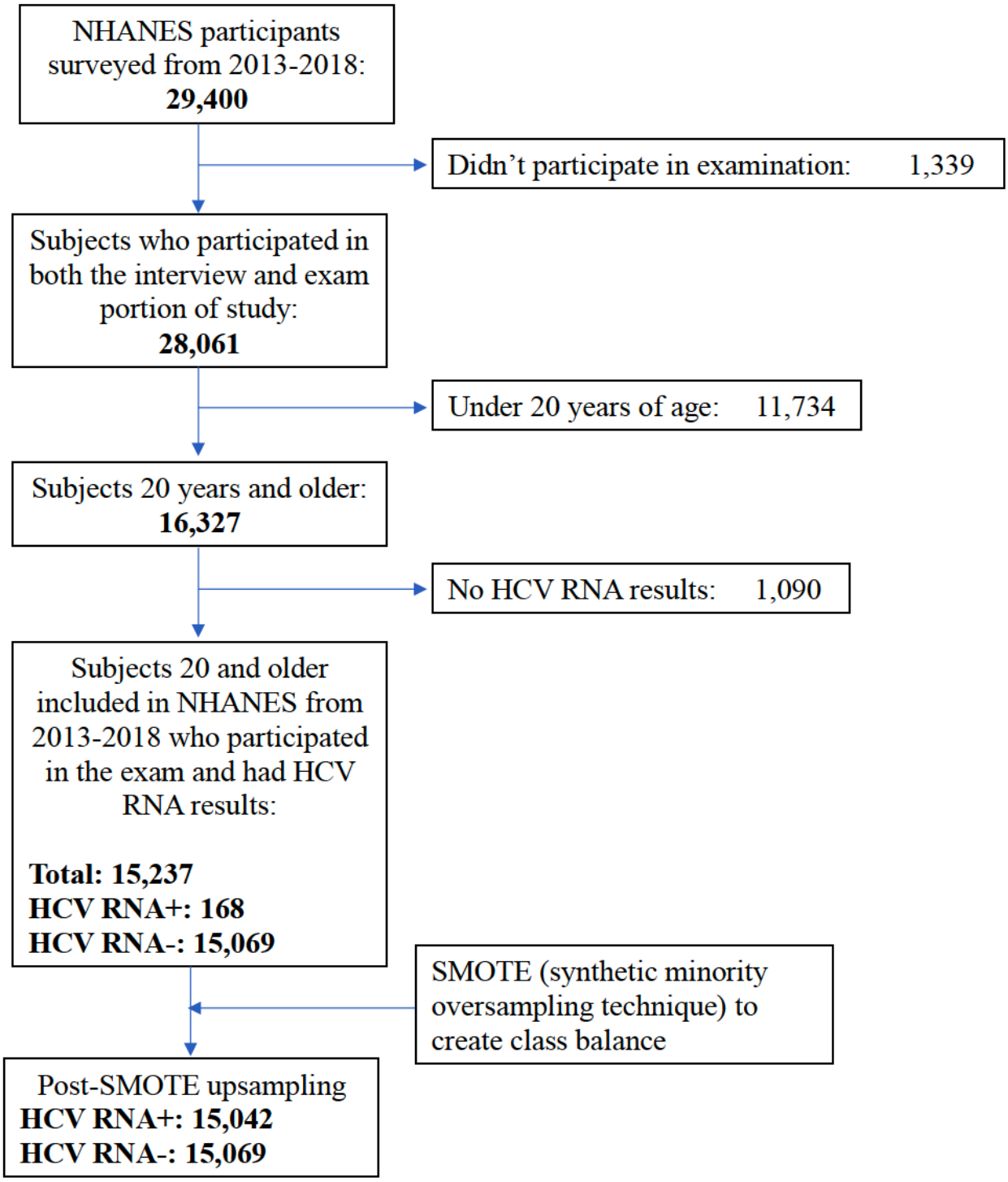


Figure 4.2: Feature engineering

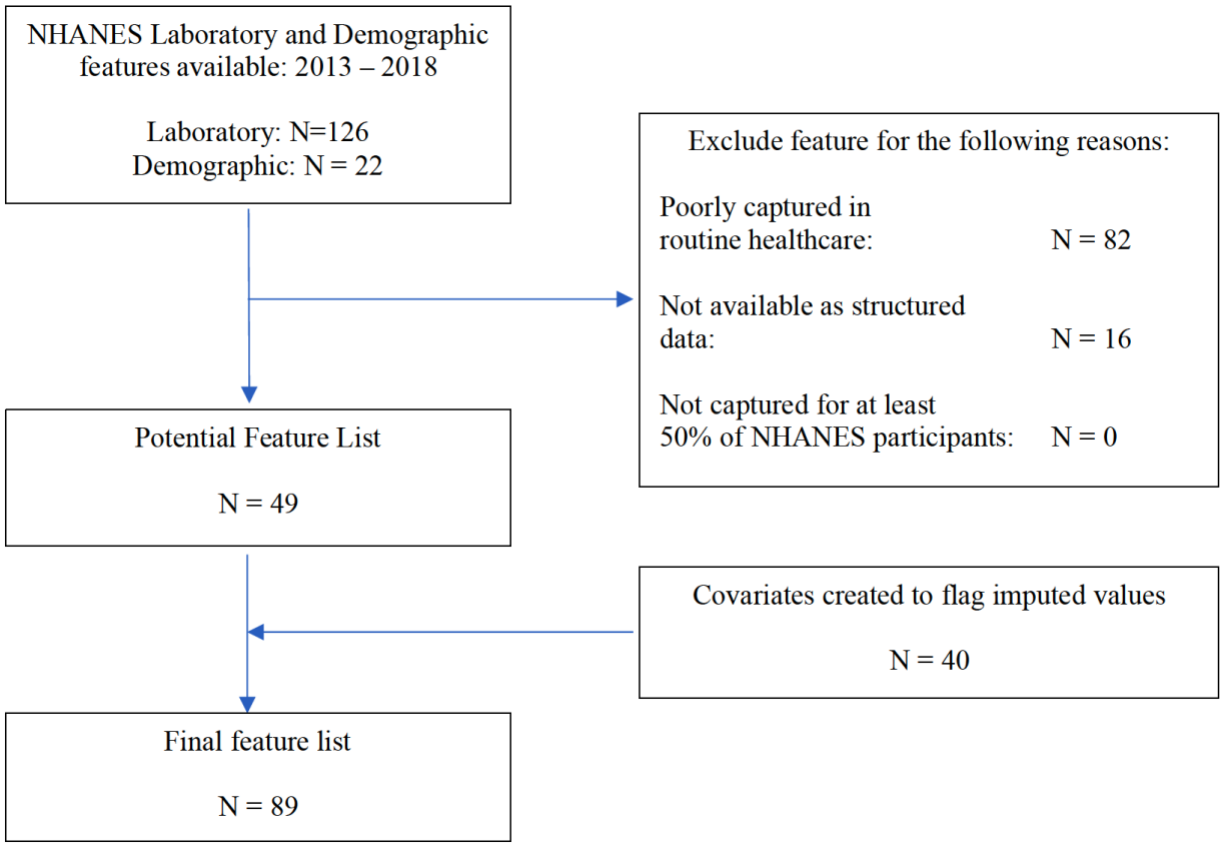
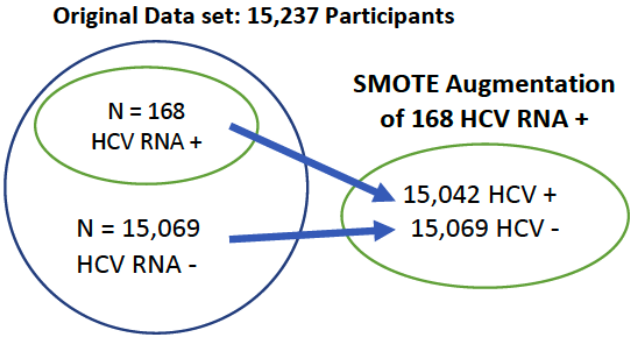


Figure 4.3: Model building

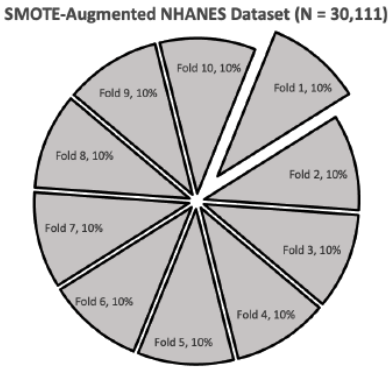
Step 1: SMOTE Augmentation

- Select all people who test positive for HCV RNA (N = 168) and augment via SMOTE (N = 15,042)
- Select all individuals who test negative for HCV RNA (N = 15,069)
- Final dataset **N = 30,111**



Step 2: Create Folds for Cross Validation

- Divide 30,111 rows into 10 folds
- Stratify folds by HCV infection status for balance



Step 3: Build Matrix of Out-Of-Sample Predictions

- Build a model for each individual learner on nine folds, predict on the 10th
- Repeat this process leaving out a different fold each time
- Build a matrix with
  - rows representing each learner
  - columns representing predictions

SMOTE-Augmented NHANES Dataset (N = 30,111)

Individual Learners	Fold1 (Build Model: Folds 2-10)	Fold2 (Build Model: Folds 1, 3-10)	Fold3 (Build Model: Folds 1-2, 4-10)	... etc...
XGBoost	Predict:XGBoost	Predict:XGBoost	Predict:XGBoost	...
Ridge	Predict:Ridge	Predict:Ridge	Predict:Ridge	...
LASSO	Predict:LASSO	Predict:LASSO	Predict:LASSO	...
GLM	Predict:GLM	Predict:GLM	Predict:GLM	...
etc...	...	...	...	...

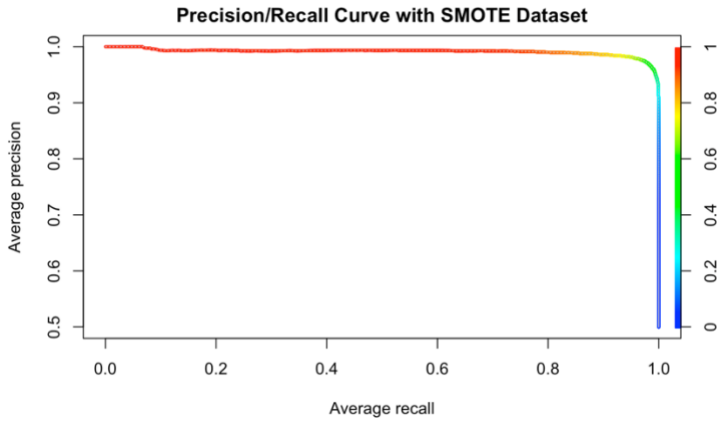
Step 4: Feed Predictions to Super Learner

- Super Learner uses predictions in the matrix to create a weighted combination that performs at least as well as each individual learner

Figure 4.4: Precision recall results for SMOTE dataset and full weighted NHANES dataset

SMOTE data set

Recall Level	Precision
0.1	0.9934
0.2	0.9937
0.3	0.9923
0.4	0.9934
0.5	0.9934
0.6	0.9932
0.7	0.9922
0.8	0.9900
0.9	0.9862



Full Weighted NHANES Dataset

Recall Level	Precision
0.1	0.5667
0.2	0.5965
0.3	0.5604
0.4	0.5113
0.5	0.5000
0.6	0.4353
0.7	0.3856
0.8	0.2885
0.9	0.1786

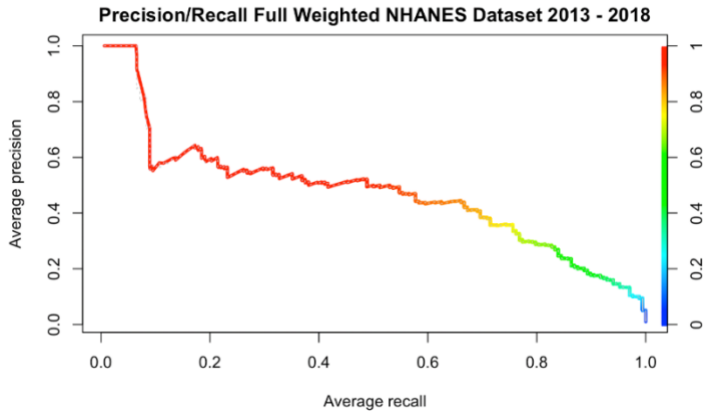


Table 4.1: Descriptive characteristics of NHANES 2013 – 2018 cohort stratified by HCV infection status

	HCV-uninfected N (%)	HCV-infected N (%)	p-value
N	**15,069 (98.9)**	**168 (1.1)**	
Demographic characteristics			
Age in years: mean (SD)	50.0 (17.7)	56.7 (10.8)	< 0.001
Male sex	7,175 (47.6)	119 (70.8)	< 0.001
Race/Ethnicity			< 0.001
Non-Hispanic White	5,666 (37.6)	59 (35.1)	
Hispanic	3,861 (25.6)	32 (19.0)	
Non-Hispanic Black	3,116 (20.7)	64 (38.1)	
Non-Hispanic Asian	1,848 (12.3)	5 (3.0)	
Other including multi-racial	578 (3.8)	8 (4.8)	< 0.001
Married	9,055 (60.1)	65 (38.7)	< 0.001
College educated	8,411 (55.9)	55 (32.7)	< 0.001
Living at or below poverty level	2,842 (21.0)	76 (49.4)	< 0.001
US born	10,330 (68.6)	148 (88.1)	< 0.001
Important laboratory measures			
Presence of hepatitis B core antibody	1,171 (7.8)	68 (40.5)	< 0.001
HIV positive	48 (0.5)	2 (2.2)	0.12
Globulin (g/L): mean (SD)	29.3 (4.5)	34.6 (6.5)	< 0.001
Albumin (ug/mL): mean (SD)	48.4 (328.0)	152.0 (621.7)	< 0.001
AST (IU/L): mean (SD)	24.1 (15.9)	53.9 (35.8)	< 0.001
ALT (IU/L): mean (SD)	23.9 (17.2)	54.81 (40.6)	< 0.001

Table 4.2: Area under the precision recall curve for individual learners and Super Learner employing a synthetic minority oversampling technique

	Original fitted Super Learner				Cross validated Super Learner			
Library of learners	**coefficients**	**AUCPR**	**SE**	**Fold SD**	**coefficients**	**AUCPR**	**SE**	**Fold SD**
GLM	0.0909	0.9986	0.0004	0.0013	0.0909	0.9984	0.0004	0.0013
BayesGLM	0.0909	0.9986	0.0004	0.0013	0.0909	0.9984	0.0004	0.0013
Ranger	0.0909	1.0000	0.0000	0.0000	0.0909	1.0000	0.0000	0.0000
GLMNet, Alpha = 0.0 (ridge)	0.0909	0.9987	0.0003	0.0009	0.0909	0.9987	0.0003	0.0009
GLMNet, Alpha = 0.2	0.0909	0.9987	0.0004	0.0013	0.0909	0.9985	0.0004	0.0013
GLMNet, Alpha = 0.4	0.0909	0.9987	0.0004	0.0013	0.0909	0.9985	0.0004	0.0013
GLMNet, Alpha = 0.6	0.0909	0.9987	0.0004	0.0013	0.0909	0.9985	0.0004	0.0013
GLMNet, Alpha = 0.8	0.0909	0.9987	0.0004	0.0013	0.0909	0.9985	0.0004	0.0013
GLMNet, Alpha = 1.0 (LASSO)	0.0909	0.9987	0.0004	0.0013	0.0909	0.9985	0.0004	0.0013
XGBoost	0.0909	0.9999	0.0001	0.0002	0.0909	0.9998	0.0001	0.0002
Dbarts	0.0909	0.9997	0.0001	0.0004	0.0909	0.9996	0.0001	0.0005
Super Learner	***NA***	**0.9995**	**0.0001**	**0.0004**	***NA***	**0.9995**	**0.0001**	**0.0005**

Table 4.3: Performance metrics on unweighted synthetic minority oversampling technique-augmented dataset and full weighted NHANES dataset

Metric	Unweighted SMOTE-augmented HCV=15,042 non-HCV =15,069	Weighted full data set HCV = 168 non-HCV = 15,069
Baseline (HCV prevalence)	0.500	0.011
Prediction threshold to maximize F1*	0.513	0.859
Maximum F1 score	0.976	0.531
Precision	1.000	0.218
Recall	0.508	0.863
Accuracy	0.753	0.964
Specificity	0.997	0.966
AUC	0.994	0.985
AUC-PR	0.991	0.477

* Threshold values from 0.001 to 0.999 were tested to see which value maximized the F1 score.

4.8: Supplemental Material

Supplemental Table S4.1: Area under the precision recall curve for all sampling methods

Sampling method	SMOTE*	ROSE**	Case control	Full dataset
Sampling method description	up-sample cases (synthetic data), all non-cases	up-sample cases (synthetic data), down-sample non-cases	all cases, random selection of non-cases 5:1	all cases, all non-cases
HCV prevalence	50.0%	50.0%	16.7%	1.1%
HCV-infected count, HCV-uninfected count	case: 15,042 non-case: 15,069	case: 7,687 non-case: 7,550	case: 168 non-case: 840	cases: 168 non-cases: 15,069
AUC-PR: Initial SL Fit	0.9995	1.0000	0.8395	0.4075
AUC-PR: Cross-validation	0.9995	1.0000	0.8424	0.4183
AUC-PR: Full data prediction	0.4609	0.0141	0.3651	0.6550

Abbreviations: SMOTE, synthetic minority oversampling technique; ROSE, random over-sampling examples, AUC-PR, area under the precision recall curve.

Chapter 5: Conclusion

In the United States, an estimated 50% of individuals infected with HCV are undiagnosed. Examining known risk factors highlights the fact that HCV infection is common in populations that often have limited access to both treatment and preventative measures that could reduce the risk of infection and reinfection. The WHO has challenged public health agencies to prioritize HCV elimination by 2030 with ambitious targets, including diagnosing 90% of individuals infected with HCV and treating 80% of the treatment-eligible population to achieve an 80% reduction in the incidence of HCV infection, and a 65% reduction in liver-related deaths[143]. To achieve these goals in the US, it is critical that policies include outreach to populations that traditionally experience obstacles in accessing healthcare.

As we strive to identify the HCV-infected population in the US and connect them with treatment, education, and strategies to avoid reinfection, there is an urgent need to explore new avenues and develop new tools to address the persistent gap in diagnosis. A review by Razavi et al. examining the timing of HCV elimination reports the alarming finding that the majority of high income countries are not making progress on elimination targets, with 30 out of 45 countries expected to be at least 20 years late in achieving the 2030 elimination goal[144]. In addition, a study by Hoenigl et al. reported a greater than 30% drop in HCV testing and treatment in the US at the start of the Covid-19 pandemic and found that treatment rates had not yet recovered from this decline[145]. A modeling study by Blach et al. estimated that each year of delay in the elimination of HCV could result in at least 44,000 additional liver cancers and more than 77,000 additional deaths globally[146].

This dissertation described the known landscape of HCV infection in the US and suggests tools and strategies for closing the gap in diagnosis. Chapter two summarized conditions under which the prevalence of HCV infection is elevated in the US. Knowledge of high-risk behaviors and conditions provides the opportunity to design policies and programs that directly target these groups.

The third chapter described the use of ensemble learning methods to identify characteristics in NHANES data that were most influential in predicting HCV infection. Because all NHANES medical examination participants were tested for HCV RNA, these data provide an opportunity to learn about individuals who were unknowingly infected with HCV. Features identified by the Super Learner algorithm as highly predictive of HCV infection included known factors, such as ALT and AST levels, and injection drug use, but also included less obvious characteristics, such as those related to cardiovascular disease, smoking, and poor dental health.

The fourth chapter described an analysis that used a subset of the NHANES data – specifically laboratory values and basic demographic information likely to be available in electronic medical records – to build an HCV infection prediction algorithm that could be used to prioritize candidates for HCV screening. The AUC-PR of the developed prediction algorithm (46.1%) represented a 42-fold improvement of the AUC-PR associated with universal screening (1.1%), suggesting that prediction algorithms such as this one could be used to optimize screening resources and increase diagnoses of HCV infection. Earlier diagnoses will enable earlier clinical interventions, resulting in improved outcomes.

Research has indicated that the existing approaches to identifying and treating individuals infected with HCV in the US are unlikely to be sufficient to achieve the WHO hepatitis elimination targets by 2030, and the result will be the continuation of preventable morbidity and mortality. It is necessary to increase the priority of the US hepatitis response and to take advantage of novel approaches and technologies, such as those described in this dissertation.

References

1. Organization, W.H. *WHO Hepatitis C Fact Sheet* 2019 February 23, 2020]; Available from: https://www.who.int/en/news-room/fact-sheets/detail/hepatitis-c.
2. El-Serag, H.B., *Hepatocellular carcinoma.* N Engl J Med, 2011. **365**(12): p. 1118-27.
3. Kanwal, F., et al., *Risk of Hepatocellular Cancer in HCV Patients Treated With Direct-Acting Antiviral Agents.* Gastroenterology, 2017. **153**(4): p. 996-1005 e1.
4. CDC. *Viral Hepatitis Surveillance Report 2018 — Hepatitis C*. 2018 [cited 2020 November 22, 2020]; Available from: https://www.cdc.gov/hepatitis/statistics/2018surveillance/HepC.htm.
5. Walker, C., *Field's Virology*. 7th ed. Vol. 1. 2020, Philadelphia, PA: Lippincott, Wiliams, & Wilkins.
6. Benova, L., et al., *Vertical transmission of hepatitis C virus: systematic review and meta-analysis.* Clin Infect Dis, 2014. **59**(6): p. 765-73.
7. Cohen, J., *Calling all baby boomers: get your hepatitis C test.* Science, 2012. **337**(6097): p. 903.
8. Prevention, C.f.D.C.a., *Recommendations for the Identification of Chronic Hepatitis C Virus Infection Among Persons Born During 1945–1965.* 2012. **61**(4).
9. Zibbell, J.E., et al., *Increases in hepatitis C virus infection related to injection drug use among persons aged </=30 years - Kentucky, Tennessee, Virginia, and West Virginia, 2006-2012.* MMWR Morb Mortal Wkly Rep, 2015. **64**(17): p. 453-8.
10. Micallef, J.M., J.M. Kaldor, and G.J. Dore, *Spontaneous viral clearance following acute hepatitis C infection: a systematic review of longitudinal studies.* J Viral Hepat, 2006. **13**(1): p. 34-41.
11. Westbrook, R.H. and G. Dusheiko, *Natural history of hepatitis C.* J Hepatol, 2014. **61**(1 Suppl): p. S58-68.
12. Alter, H.J., *HCV natural history: the retrospective and prospective in perspective.* J Hepatol, 2005. **43**(4): p. 550-2.
13. El-Serag, H., *Hepatocellular Carcinoma.* New England Journal of Medicine, 2011. **365**: p. 1118-27.
14. Manns, M.P., H. Wedemeyer, and M. Cornberg, *Treating viral hepatitis C: efficacy, side effects, and complications.* Gut, 2006. **55**(9): p. 1350-9.
15. Afdhal, N., et al., *Ledipasvir and sofosbuvir for untreated HCV genotype 1 infection.* N Engl J Med, 2014. **370**(20): p. 1889-98.
16. Afdhal, N., et al., *Ledipasvir and sofosbuvir for previously treated HCV genotype 1 infection.* N Engl J Med, 2014. **370**(16): p. 1483-93.
17. WHO, *Progress Report on Access to Hepatitis C Treatment; Focus on Overcoming Barriers in Low- and Middle-Income Countries*. 2018.
18. Hagan, H., et al., *Self-reported hepatitis C virus antibody status and risk behavior in young injectors.* Public Health Rep, 2006. **121**(6): p. 710-9.
19. Zhou, K. and N.A. Terrault, *Gaps in Viral Hepatitis Awareness in the United States in a Population-based Study.* Clin Gastroenterol Hepatol, 2020. **18**(1): p. 188-195 e4.
20. Klevens, R.M., et al., *Estimating acute viral hepatitis infections from nationally reported cases.* Am J Public Health, 2014. **104**(3): p. 482-7.
21. Onofrey, S., et al., *Underascertainment of acute hepatitis C virus infections in the U.S. surveillance system: a case series and chart review.* Ann Intern Med, 2015. **163**(4): p. 254-61.
22. Colvin, H.M. and A.E. Mitchell, *Hepatitis and Liver Cancer: A National Strategy for Prevention and Control of Hepatitis B and C*. 2010, National Academies Press (US): Washington (DC).
23. El-Serag, H., A.S. Lok, and D.L. Thomas, *The dawn of a new era: transforming our domestic response to hepatitis B & C.* Gastroenterology, 2010. **138**(4): p. 1225-30, 1230 e1-3.
24. USPSTF, *Screening for Hepatitis C Virus Infection in Adults: Recommendation Statement.* Annals of Internal Medicine, 2004. **140**(6): p. 462-464.
25. Szabo, E., et al., *Viral hepatitis: new data on hepatitis C infection.* Pathol Oncol Res, 2003. **9**(4): p. 215-21.
26. Lingala, S. and M.G. Ghany, *Natural History of Hepatitis C.* Gastroenterol Clin North Am, 2015. **44**(4): p. 717-34.
27. USPSTF. *Procedure Manual Appendix I. Congressional Mandate Establishing the U.S. Preventive Services Task Force*. 2019; Available from: https://www.uspreventiveservicestaskforce.org/uspstf/procedure-manual/procedure-manual-appendix-i.

28. Denniston, M.M., et al., *Chronic hepatitis C virus infection in the United States, National Health and Nutrition Examination Survey 2003 to 2010.* Ann Intern Med, 2014. **160**(5): p. 293-300.
29. Force, U.S.P.S.T., *Screening for Hepatitis C Virus Infection in Adults: U.S. Preventive Services Task Force Recommendation Statement.* Annals of Internal Medicine, 2013. **159**(5): p. 349-58.
30. Barocas, J.A., et al., *Hepatitis C Testing Increased Among Baby Boomers Following The 2012 Change To CDC Testing Recommendations.* Health Aff (Millwood), 2017. **36**(12): p. 2142-2150.
31. Buckley, G.J. and B.L. Strom, *A National Strategy for the Elimination of Viral Hepatitis Emphasizes Prevention, Screening, and Universal Treatment of Hepatitis C.* Ann Intern Med, 2017. **166**(12): p. 895-896.
32. European Union HCV Collaborators, *Hepatitis C virus prevalence and level of intervention required to achieve the WHO targets for elimination in the European Union by 2030: a modelling study.* Lancet Gastroenterol Hepatol, 2017. **2**(5): p. 325-336.
33. Kabiri, M., et al., *The changing burden of hepatitis C virus infection in the United States: model-based predictions.* Ann Intern Med, 2014. **161**(3): p. 170-80.
34. USPSTF. *Hepatitis C Virus Infection in Adolescents and Adults: Screening*. 2020 [cited 2020 November 11, 2020]; Screening recommendations of hepatitis C]. Available from: https://www.uspreventiveservicestaskforce.org/uspstf/recommendation/hepatitis-c-screening.
35. Koretz, R.L., et al., *Is widespread screening for hepatitis C justified?* BMJ, 2015. **350**: p. g7809.
36. Winetsky, D., et al., *Attitudes, practices and perceived barriers to hepatitis C screening among medical residents at a large urban academic medical center.* J Viral Hepat, 2019. **26**(11): p. 1355-1358.
37. Radwan, D., et al., *HCV Screening and Treatment Uptake Among Patients in HIV Care During 2014-2015.* J Acquir Immune Defic Syndr, 2019. **80**(5): p. 559-567.
38. Shehata, N., et al., *Barriers to and facilitators of hepatitis C virus screening and testing: A scoping review.* Can Commun Dis Rep, 2018. **44**(7-8): p. 166-172.
39. McCauley, M., et al., *Screening Adult Children of Hepatitis C-Infected Baby Boomers: Barriers to Testing and Prevalence Estimates.* Clin Liver Dis (Hoboken), 2020. **16**(2): p. 77-82.
40. Alter, H.J. and F.V. Chisari, *Is Elimination of Hepatitis B and C a Pipe Dream or Reality?* Gastroenterology, 2019. **156**(2): p. 294-296.
41. Chhatwal, J. and N.L. Sussman, *Universal Screening for Hepatitis C: An Important Step in Virus Elimination.* Clin Gastroenterol Hepatol, 2019. **17**(5): p. 835-837.
42. Eckman, M.H., J.W. Ward, and K.E. Sherman, *Cost Effectiveness of Universal Screening for Hepatitis C Virus Infection in the Era of Direct-Acting, Pangenotypic Treatment Regimens.* Clin Gastroenterol Hepatol, 2019. **17**(5): p. 930-939 e9.
43. Linas, B.P., et al., *Cost Effectiveness and Cost Containment in the Era of Interferon-Free Therapies to Treat Hepatitis C Virus Genotype 1.* Open Forum Infect Dis, 2017. **4**(1): p. ofw266.
44. Barocas, J.A., et al., *Population-level Outcomes and Cost-Effectiveness of Expanding the Recommendation for Age-based Hepatitis C Testing in the United States.* Clin Infect Dis, 2018. **67**(4): p. 549-556.
45. Scott, N., et al., *A model of the economic benefits of global hepatitis C elimination: an investment case.* Lancet Gastroenterol Hepatol, 2020. **5**(10): p. 940-947.
46. de Quadros, C.A., et al., *Measles eradication: experience in the Americas.* Bulletin of the World Health Organization, 1998. **76 Suppl 2**: p. 47-52.
47. Fenner, F., *Candidate viral diseases for elimination or eradication.* Bull World Health Organ, 1998. **76 Suppl 2**: p. 68-70.
48. Henderson, D.A., *Eradication: lessons from the past.* Bulletin of the World Health Organization, 1998. **76 Suppl 2**: p. 17-21.
49. Ward, J.W. and A.R. Hinman, *What Is Needed to Eliminate Hepatitis B Virus and Hepatitis C Virus as Global Health Threats.* Gastroenterology, 2019. **156**(2): p. 297-310.
50. Drainoni, M.L., et al., *Effectiveness of a risk screener in identifying hepatitis C virus in a primary care setting.* Am J Public Health, 2012. **102**(11): p. e115-21.
51. Trinh, J. and N. Turner, *Improving adherence to hepatitis C screening guidelines.* BMJ Open Qual, 2018. **7**(2): p. e000108.
52. Hojat, L., et al., *Doubling Hepatitis C Virus Screening in Primary Care Using Advanced Electronic Health Record Tools-A Non-Randomized Controlled Trial.* J Gen Intern Med, 2020. **35**(2): p. 498-504.
53. Dieterich, D.T., *A Simplified Algorithm for the Management of Hepatitis C Infection.* Gastroenterol Hepatol (N Y), 2019. **15**(5 Suppl 3): p. 1-12.

54. Prevention, C.f.D.C.a. *Hepatitis C Questions and Answers for Health Professionals*. 2022 February 11, 2022]; Available from: https://www.cdc.gov/hepatitis/hcv/hcvfaq.htm#b1.
55. Gilead Sciences, I. *U.S. Food and Drug Administration Approves Gilead's Sovaldi™ (Sofosbuvir) for the Treatment of Chronic Hepatitis C* 2013 [cited 2022 Feb 28, 2022]; Available from: https://www.gilead.com/news-and-press/press-room/press-releases/2013/12/us-food-and-drug-administration-approves-gileads-sovaldi-sofosbuvir-for-the-treatment-of-chronic-hepatitis-c.
56. Organization, W.H., *Global Health Sector Strategy on Viral Hepatitis, 2016–2021*. 2016: Geneva.
57. Hofmeister, M.G., et al., *Estimating Prevalence of Hepatitis C Virus Infection in the United States, 2013-2016.* Hepatology, 2019. **69**(3): p. 1020-1031.
58. Prevention, C.f.D.C.a., *Viral Hepatitis Surveillance, United States, 2016*. 2018.
59. Prevention, C.f.D.C.a. *Viral Hepatitis Surveillance Report, 2019*. 2019 [cited 2022 February 11, 2022]; Available from: https://www.cdc.gov/hepatitis/statistics/2019surveillance/index.htm.
60. Terrault, N.A., *Hepatitis C elimination: challenges with under-diagnosis and under-treatment[version 1; referees: 2 approved]*. F1000 Research, 2019. **8(F1000 Faculty Rev)**(54): p. 1-11.
61. Spearman, C.W., et al., *Hepatitis C.* The Lancet, 2019. **394**(10207): p. 1451-1466.
62. Schillie, S., et al., *CDC Recommendations for Hepatitit C screening Among Adults - United States, 2020.* MMWR Morb Mortal Wkly Rep, 2020. **69 (i, RR-2)**: p. 1-17.
63. Cooke, G.S., et al., *Accelerating the elimination of viral hepatitis: a Lancet Gastroenterology & Hepatology Commission.* Lancet Gastroenterol Hepatol, 2019. **4**(2): p. 135-184.
64. Gamkrelidze, I., et al., *Progress towards hepatitis C virus elimination in high-income countries: An updated analysis.* Liver Int, 2021. **41**(3): p. 456-463.
65. Larney, S., et al., *Incidence and prevalence of hepatitis C in prisons and other closed settings: results of a systematic review and meta-analysis.* Hepatology, 2013. **58**(4): p. 1215-24.
66. Jordan, A.E., et al., *Prevalence of hepatitis C virus infection among HIV+ men who have sex with men: a systematic review and meta-analysis.* International Journal of STD & AIDS, 2017. **28**(2): p. 145-159.
67. Wirtz, A.L., et al., *HIV and Viral Hepatitis Among Imprisoned Key Populations.* Epidemiologic Reviews, 2018. **40**(1): p. 12-26.
68. Bruce, V., et al., *Hepatitis C Virus Infection in Indigenous Populations in the United States and Canada.* Epidemiologic Reviews, 2019. **41**(1): p. 158-167.
69. Zhou, B., et al., *Factors Correlating to the Development of Hepatitis C Virus Infection among Drug Users-Findings from a Systematic Review and Meta-Analysis.* International Journal of Environmental Research & Public Health [Electronic Resource], 2019. **16**(13): p. 02.
70. Schalkoff, C.A., et al., *The opioid and related drug epidemics in rural Appalachia: A systematic review of populations affected, risk factors, and infectious diseases.* Substance Abuse, 2020. **41**(1): p. 35-69.
71. Arum, C., et al., *Homelessness, unstable housing, and risk of HIV and hepatitis C virus acquisition among people who inject drugs: a systematic review and meta-analysis.* The lancet. Public Health, 2021. **6**(5): p. e309-e323.
72. Jin, F., et al., *Prevalence and incidence of hepatitis C virus infection in men who have sex with men: a systematic review and meta-analysis.* The Lancet. Gastroenterology & Hepatology, 2021. **6**(1): p. 39-56.
73. Fill, M.A., et al., *Epidemiology and risk factors for hepatitis C virus infection in a high-prevalence population.* Epidemiol Infect, 2018. **146**(4): p. 508-514.
74. Keen, L., 2nd, et al., *Injection and non-injection drug use and infectious disease in Baltimore City: differences by race.* Addict Behav, 2014. **39**(9): p. 1325-8.
75. Ayano, G., et al., *A systematic review and meta-analysis of gender difference in epidemiology of HIV, hepatitis B, and hepatitis C infections in people with severe mental illness.* Annals of General Psychiatry, 2018. **17**: p. 16.
76. Armstrong, G., et al., *The Prevalence of Hepatitis C Virus Infecion in the United States, 1999 through 2002.* Annals of Internal Medicine, 2006. **144**: p. 705 - 715.
77. Hughes, E., et al., *Prevalence of HIV, hepatitis B, and hepatitis C in people with severe mental illness: a systematic review and meta-analysis.* The Lancet. Psychiatry, 2016. **3**(1): p. 40-48.
78. Gower, E., et al., *Global prevalence and genotype distribution of hepatitis C virus infection in 2015: a modelling study.* Lancet Gastroenterol Hepatol, 2017. **2**(3): p. 161-176.
79. Westermann, C., et al., *The prevalence of hepatitis C among healthcare workers: a systematic review and meta-analysis.* Occupational & Environmental Medicine, 2015. **72**(12): p. 880-8.
80. Dolan, K., et al., *Global burden of HIV, viral hepatitis, and tuberculosis in prisoners and detainees.* Lancet, 2016. **388**: p. 1089 - 1102.

81. Beijer, U., A. Wolf, and S. Fazel, *Prevalence of tuberculosis, hepatitis C virus, and HIV in homeless people: a systematic review and meta-analysis.* The Lancet Infectious Diseases, 2012. **12**(11): p. 859-70.
82. Choo, Q.L., et al., *Isolation of a cDNA Clone Derived from a Blood-Borne Non-A, Non-B Viral Hepatitis Genome*. 1989: Science. p. 359 - 362.
83. Houghton, M., *Discovery of the hepatitis C virus.* Liver Int, 2009. **29 Suppl 1**: p. 82-8.
84. Jorgennsen, C., *"Know More Hepatitis:" CDC's National Education Campaign to Increase Hepatitis C Testing Among People Born Between 1945 and 1965.* Public Health Reports, 2016. **131**: p. 29 - 34.
85. Hines, L.A., et al., *Associations between national development indicators and the age profile of people who inject drugs: results from a global systematic review and meta-analysis.* The Lancet Global Health, 2020. **8**(1): p. e76-e91.
86. Ryerson, A., et al., *Newly Reported Acute and Chronic Hepatitis C Cases — United States*, in *Vital Signs*, M.M.M.W. Rep, Editor. 2020. p. 399–404.
87. Scheinmann, R., et al., *Non-Injection Drug Use and Hepatitis C Virus: A Systematic Review.* Drug and Alcohol Dependence, 2007. **89**(1-12).
88. Degenhardt, L., et al., *Global prevalence of injecting drug use and sociodemographic characteristics and prevalence of HIV, HBV, and HCV in people who inject drugs: a multistage systematic review.* The Lancet Global Health, 2017. **5**(12): p. e1192-e1207.
89. Shiffman, M.L., *The next wave of hepatitis C virus: The epidemic of intravenous drug use.* Liver Int, 2018. **38 Suppl 1**: p. 34-39.
90. Prevention, C.f.D.C.a. *Infection Prevention during Blood Glucose Monitoring and Insulin Administration*. Injection Safety 2011; Available from: https://www.cdc.gov/injectionsafety/blood-glucose-monitoring.html.
91. Prevention, C.f.D.C.a. *Blood Safety Basics*. 2020; Available from: https://www.cdc.gov/bloodsafety/basics.html.
92. Pena-Orellana, M., et al., *Prevalence of HCV risk behaviors among prison inmates: tattooing and injection drug use.* J Health Care Poor Underserved, 2011. **22**(3): p. 962-82.
93. Lin, O.N., et al., *HCV Prevalence in Asian Americans in California.* J Immigr Minor Health, 2017. **19**(1): p. 91-97.
94. Tohme, R.A. and S.D. Holmberg, *Transmission of hepatitis C virus infection through tattooing and piercing: a critical review.* Clin Infect Dis, 2012. **54**(8): p. 1167-78.
95. Patel, P.R., et al., *Epidemiology, surveillance, and prevention of hepatitis C virus infections in hemodialysis patients.* Am J Kidney Dis, 2010. **56**(2): p. 371-8.
96. Prasad, M.R., *Hepatitis C Virus Screening in Pregnancy: Is It Time to Change Our Practice?* Obstet Gynecol, 2016. **128**(2): p. 229-230.
97. Rosenberg, E.S., et al., *Prevalence of Hepatitis C Virus Infection in US States and the District of Columbia, 2013 to 2016.* JAMA Netw Open, 2018. **1**(8): p. e186371.
98. Chou, R., et al., *Screening for Hepatitis C Virus Infection in Adolescents and Adults: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force.* JAMA, 2020.
99. Di Bisceglie, A.M., *Natural history of hepatitis C: its impact on clinical management.* Hepatology, 2000. **31**(4): p. 1014-8.
100. Grebely, J., et al., *Hepatitis C virus clearance, reinfection, and persistence, with insights from studies of injecting drug users: towards a vaccine.* Lancet Infect Dis, 2012. **12**(5): p. 408-14.
101. Chen, L.P., et al., *Antiviral treatment to prevent chronic hepatitis B or C-related hepatocellular carcinoma.* World J Virol, 2012. **1**(6): p. 174-83.
102. Denniston, M.M., et al., *Awareness of infection, knowledge of hepatitis C, and medical follow-up among individuals testing positive for hepatitis C: National Health and Nutrition Examination Survey 2001-2008.* Hepatology, 2012. **55**(6): p. 1652-61.
103. Schillie, S., et al., *CDC Recommendations for Hepatitis C Screening Among Adults - United States, 2020.* MMWR Recomm Rep, 2020. **69**(2): p. 1-17.
104. Doyle, O.M., N. Leavitt, and J.A. Rigg, *Finding undiagnosed patients with hepatitis C infection: an application of artificial intelligence to patient claims data.* Sci Rep, 2020. **10**(1): p. 10521.
105. Dinh, A., et al., *A data-driven approach to predicting diabetes and cardiovascular disease with machine learning.* BMC Med Inform Decis Mak, 2019. **19**(1): p. 211.
106. Oh, J., et al., *Identifying depression in the National Health and Nutrition Examination Survey data using a deep learning algorithm.* J Affect Disord, 2019. **257**: p. 623-631.

107. CDC, *Testing for HCV infection: an update of guidance for clinicians and laboratorians.* MMWR Morb Mortal Wkly Rep, 2013. **62**(18): p. 362-5.
108. Gelman, A., et al., *A weakly informative default prior distribution for logistic and other regression models.* The Annals of Applied Statistics, 2008. **2**(4).
109. Starkweather, J. *Bayesian Generalized Linear Models in R*. Research and Statistical Support, 2011.
110. Friedman, J.H., T.; Tibsharani, R.;, *Regularization Paths for Generalized Linear Models via Coordinate Descent.* Journal of Statistical Software, 2010. **33**(1): p. 1-22.
111. C., C.T.G., *XGBoost: A Scalable Tree Boosting System.* KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016: p. 785-794.
112. Dorie, V., *dbarts: Discrete Bayesian Additive Regression Trees Sampler.* 2020.
113. Wright, N.Z., A., *ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R.* Journal of Statistical Software, 2017. **77**(1): p. 1-17.
114. Gini, C., *The Concise Encyclopedia of Statistics*, in *The Concise Encyclopedia of Statistics*. 2008, Springer New York: New York, NY. p. 231-233.
115. Flach, P.A.K., M., *Precision-Recall-Gain Curves: PR Analysis Done Right*, in *Advances in Neural Information Processing Systems 28 (NIPS 2015)*. 2015: Montreal, CA.
116. van Rijsbergen, C.J., *Information Retrieval*, in *Information Retrieval*, Butterworth-Heinemann, Editor. 1979: Newton, MA, USA.
117. Haga, H., et al., *A machine learning-based treatment prediction model using whole genome variants of hepatitis C virus.* PLoS One, 2020. **15**(11): p. e0242028.
118. Ohrvall, M., L. Berglund, and B. Vessby, *Sagittal abdominal diameter compared with other anthropometric measurements in relation to cardiovascular risk.* International Journal of Obesity, 2000. **24**(4): p. 497-501.
119. Coates, E.A., et al., *Hepatitis C infection and associated oral health problems.* Australian Dental Journal, 2000. **45**(2): p. 108 - 114.
120. Alaizari, N.A., et al., *Hepatitis C virus infections in oral lichen planus: a systematic review and meta-analysis.* Aust Dent J, 2016. **61**(3): p. 282-7.
121. Arum, C., et al., *Homelessness, unstable housing, and risk of HIV and hepatitis C virus acquisition among people who inject drugs: a systematic review and meta-analysis.* The Lancet Public Health, 2021. **6**(5): p. e309-e323.
122. Zampino, R., et al., *Hepatitis C virus infection and prisoners: Epidemiology, outcome and treatment.* World J Hepatol, 2015. **7**(21): p. 2323-30.
123. Hack, B., et al., *Oral Prescription Opioids as a High-Risk Indicator for Hepatitis C Infection: Another Step Toward HCV Elimination.* Journal of Primary Care & Community Health, 2021. **12**: p. 1-9.
124. Miech, R., et al., *Prescription Opioids in Adolescence and Future Opioid Misuse.* Pediatrics, 2015. **136**(5): p. e1169-77.
125. El-Serag, H.B., *Epidemiology of viral hepatitis and hepatocellular carcinoma.* Gastroenterology, 2012. **142**(6): p. 1264-1273 e1.
126. Sene, D., N. Limal, and P. Cacoub, *Hepatitis C virus-associated extrahepatic manifestations: a review.* Metab Brain Dis, 2004. **19**(3-4): p. 357-81.
127. Sebastiani, G., K. Gkouvatsos, and K. Pantopoulos, *Chronic hepatitis C and liver fibrosis.* World J Gastroenterol, 2014. **20**(32): p. 11033-53.
128. Cacoub, P., et al., *Extrahepatic manifestations of chronic hepatitis C virus infection.* Ther Adv Infect Dis, 2016. **3**(1): p. 3-14.
129. CDC. *2019 HCV Surveillance*. 2019 [cited 2021 November 21, 2021]; Available from: https://www.cdc.gov/hepatitis/statistics/2019surveillance/Table3.7.htm.
130. WHO, *Global Hepatitis Report*. 2017: Geneva.
131. Fried, M.W., et al., *Peginterferon alfa-2a plus ribavirin for chronic hepatitis C virus infection.* N Engl J Med, 2002. **347**(13): p. 975-82.
132. Rong, L. and A.S. Perelson, *Treatment of hepatitis C virus infection with interferon and small molecule direct antivirals: viral kinetics and modeling.* Crit Rev Immunol, 2010. **30**(2): p. 131-48.
133. Desai, R.J., et al., *Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes.* JAMA Netw Open, 2020. **3**(1): p. e1918962.
134. Rajkomar, A., et al., *Scalable and accurate deep learning with electronic health records.* NPJ Digit Med, 2018. **1**: p. 18.

135. CDC. *Research Ethics Review Board (ERB) Approval*. 2021 [cited 2021 November 21, 2021]; Available from: https://www.cdc.gov/nchs/nhanes/irba98.htm.
136. CDC. *Hepatitis C virus RNA in Serum*. 2021 [cited 2021 November 20, 2021]; National Health and Nutrition Examination Survey 2013-2014 Data Documentation, Codebook, and Frequencies]. Available from: https://wwwn.cdc.gov/Nchs/Nhanes/2013-2014/HEPC_H.htm.
137. Chawla, N., et al., *SMOTE: Synthetic Minority Over-sampling Technique.* JAIP, 2002. **16**(1): p. 321-357.
138. Menardi, G. and N. Torelli, *Training and assessing classification rules with imbalanced data.* Data Min Knowl Disc, 2014(28): p. 92-122.
139. Polley, E.C. and M.J. van der Laan, *Super Learner in Prediction.* U.C. Berkeley Division of Biostatistics Working Paper Series, 2010.
140. LeDell, E., M.J. van der Laan, and M. Petersen, *AUC-Maximizing Ensembles through Metalearning.* Int J Biostat, 2016. **12**(1): p. 203-18.
141. Weiss, G. *Learning with Rare Cases and Small Disjuncts*. in *12th International Conference on Machine Learning*. 1995. Morgan Kaufman.
142. Orooji, A. and F. Kermani, *Machine Learning Based Methods for Handling Imbalanced Data in Hepatitis Diagnosis.* Frontiers in Health Informatics, 2021. **10**: p. 57.
143. Organization, W.H., *GLOBAL HEALTH SECTOR STRATEGY ON VIRAL HEPATITIS 2016–2021*. 2016: Geneva, Switzerland.
144. Razavi, H., et al., *Global timing of hepatitis C virus elimination in high-income countries.* Liver Int, 2020. **40**(3): p. 522-529.
145. Hoenigl, M., et al., *Sustained impact of the COVID-2019 pandemic on HCV treatment initiations in the United States* Clinical Infectious Diseases, 2022. **ciac175**.
146. Blach, S., et al., *Impact of COVID-19 on global HCV elimination efforts.* J Hepatol, 2020.