Finding undiagnosed patients with hepatitis C virus: an application of state-of-the-art machine learning methods

Orla M. Doyle[1], Harsha Jayanti[1], Daniel Homola[1], and John A. Rigg[1].

[1] Predictive Analytics, Real World Insights, IQVIA, London, N1 9JY, UK.

Email: Orla.Doyle@iqvia; John.Rigg@iqvia.com.

IMS Health & Quintiles are now IQVIA™

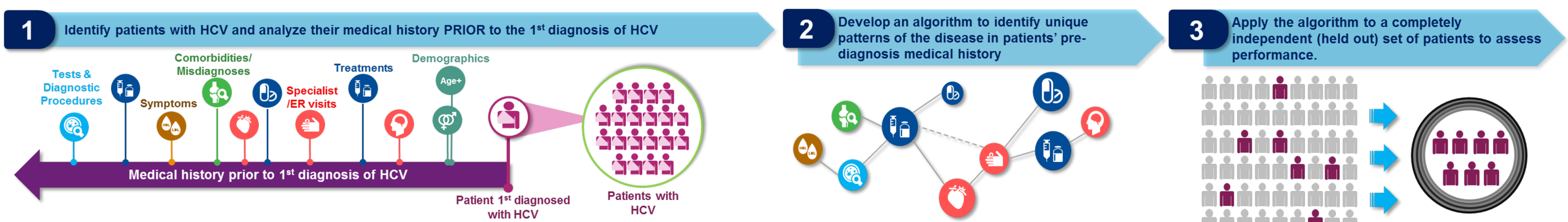

Figure 1. The pre-diagnosed digital footprint of HCV patients was used to find undiagnosed patients.

Introduction

The Hepatitis C Virus (HCV) is a chronic, life-threatening disease which is substantially under-diagnosed. Accelerating time to diagnosis can lead to earlier treatment and improved patient outcomes. This was a retrospective database study to develop an algorithm which could be used to identify undiagnosed patients with HCV based on routinely collected patient data.

Methods

Data were extracted from US prescription and open-source medical claims between 2010 and 2016. Outcomes for HCV patients were coded as 1; outcomes for non-HCV patients were set to 0. Index date for HCV patients was the first observed date of diagnosis, ensuring only pre-diagnosed attributes were used. The most recent activity was used as the index date for non-HCV patients. Features captured information on demographics, treatments, procedures and symptomatology, comorbidities/misdiagnoses and specialist visits.

Stratification criteria were implemented to ensure that the analysis focused on a broad pool of patients with some minimal level of risk of HCV. Patients were randomly assigned to a training set, a validation set and a test/hold-out set accounting for 80%, 10% and 10% of patients respectively. A case-control design was used whereby for each HCV patient (in the validation and test sets), 200 non-HCV patients were matched based on length of lookback window and timing of the index date. The ratio of cases to controls represents an estimate of the prevalence of undiagnosed HCV in the US population. Binary classifiers were estimated based on:

- **Conventional parametric methods**
  - Logistic regression (unconstrained)
- **Non-parametric machine learning methods**
  - Random forest [1],
  - Gradient boosting trees [2]
- **An ensemble of classifiers based on gradient boosted trees**
  - An ensemble classifier combines output from several individual classifiers to form what is sometimes referred to as a 'super-learner' [3]. A super-learner has the potential of capturing different properties from different methods to form a single classifier which can improve overall predictive performance.

Results

In total, approximately 10 million HCV and non-HCV patients were selected for inclusion after applying the stratification criteria. There were 160 predictors (independent variables) in total. The demographics of HCV and non-HCV patients are provided in Table 1, showing a similar distribution of age and gender by cohort

Variable	HCV	non-HCV
Patient counts	120,023	9,601,900
Age (mean +/- s.d.)	50.9+/-14.2	54.2+/-14.1
Gender	53% M, 47% F	40% M, 60% F

Table 1 Key characteristics of HCV and non-HCV patients.

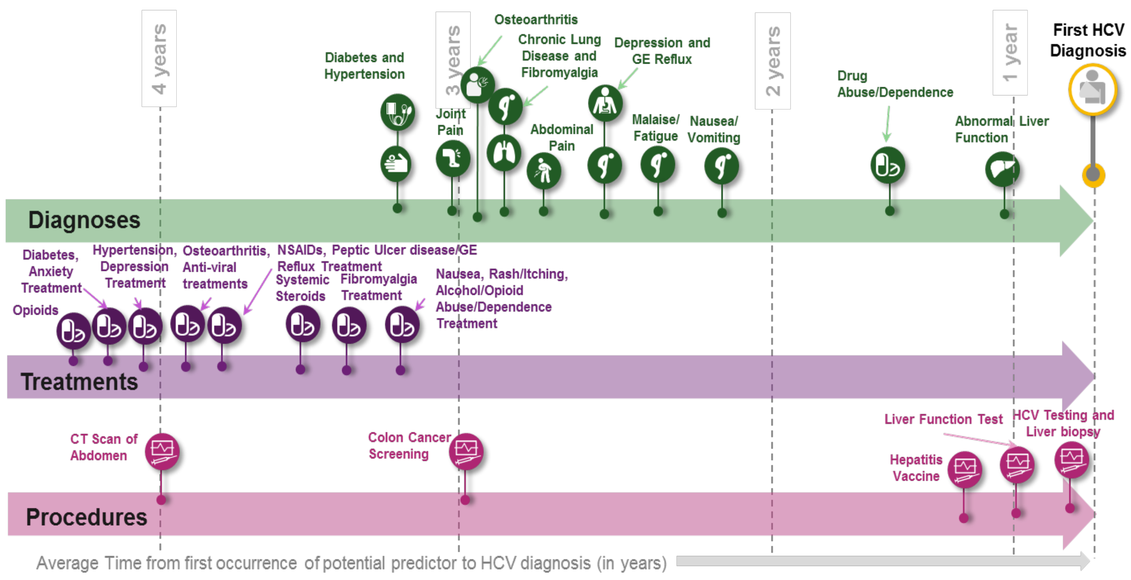

Figure 2 HCV patient medical journey prior to diagnosis timeline

In Figure 2, the HCV patient journey illustrates that patients begin experiencing known symptoms of HCV on average 2-3 years prior to their diagnosis. Treatment with NSAIDs, systemic steroids, opioids occur earlier than their diagnoses indicating that these patients are seeking treatments for these symptoms prior to receiving their diagnosis for HCV. Patients are also undergoing several diagnostic test procedures close to the time of their diagnoses.

The precision-recall curves for each model are illustrated in Figure 3. For 50% recall, precision was 98%, 87%, 4%, 2% for the ensemble, gradient boosting, random forest, logistic regression, respectively. The top 10 most important variables for the ensemble model are illustrated in Figure 4.

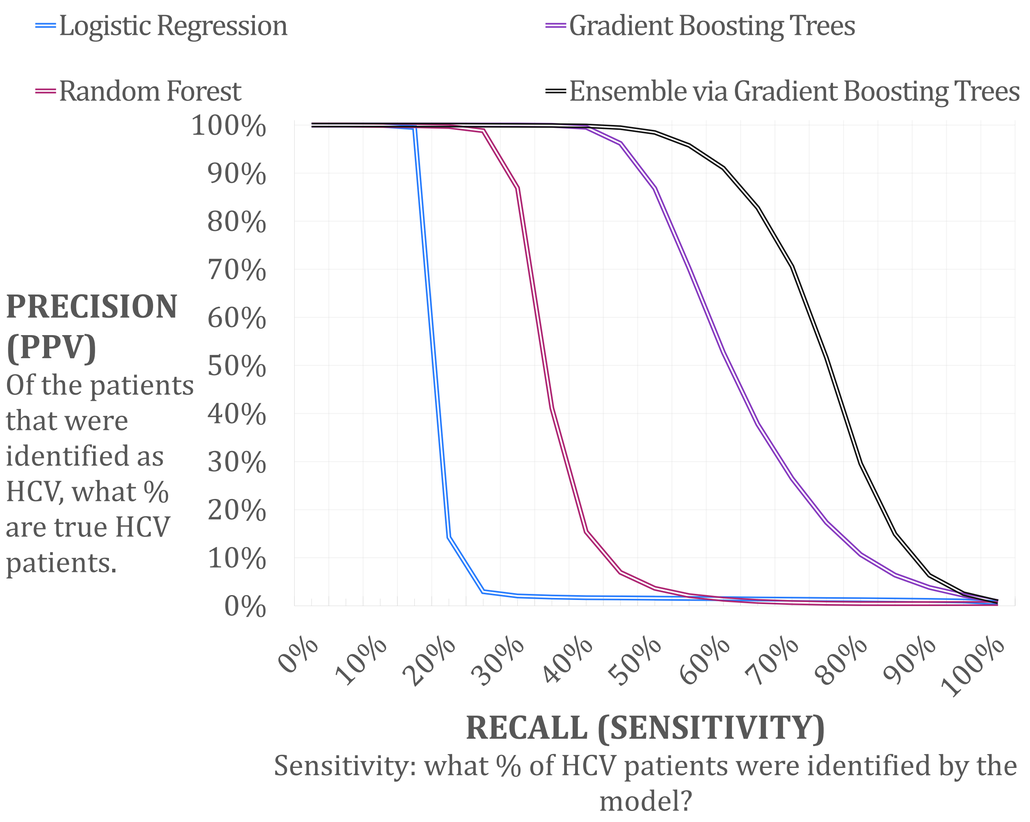

Figure 3 Precision-recall curves for different modelling approaches.

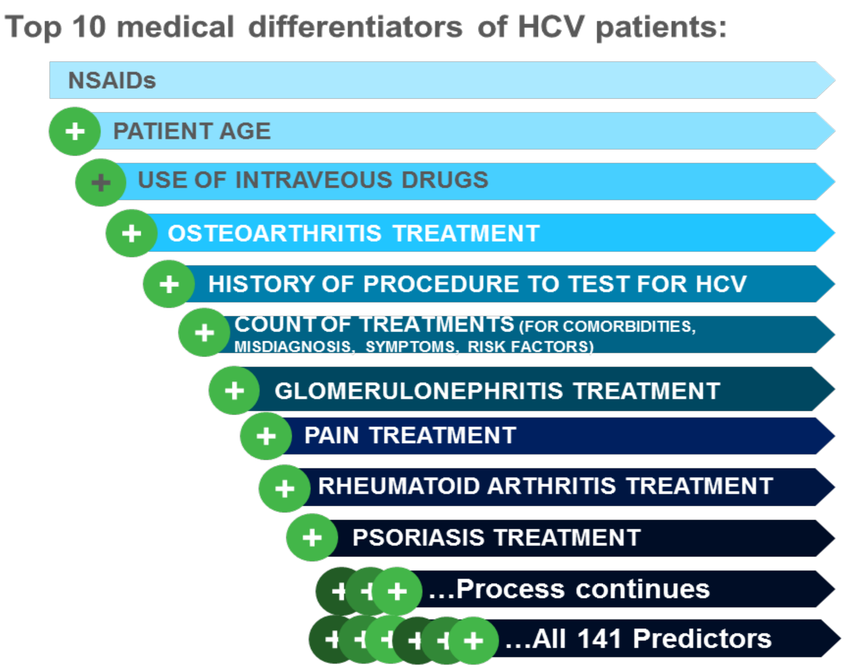

Figure 4 Variable importance for the ensemble model using Gradient Boosting Trees.

Conclusions

The evidence suggests that algorithms leveraging routinely collected real-world data could be a highly effective way to screen for undiagnosed HCV patients. Using state-of-art machine learning approaches to combine parametric and non-parametric models out-performed conventional approaches, highlighting the potential value of these methods.

References

1. L. Breiman. *Random forests*. Machine learning, 45(1):5–32, 2001
2. J. Friedman. *Greedy function approximation: a gradient boosting machine*. Annals of Statistics, 29(5):1189–1232, 2001.
3. D. Wolpert, *Stacked Generalization.*, Neural Networks, 5(2), pp. 241-259., 199

ISPOR 20th Annual European Congress, 4th - 8th Nov 2017, Glasgow, Scotland.