IMS Health & Quintiles are now IQVIA™

Using machine learning and real world data to tackle complex healthcare challenges

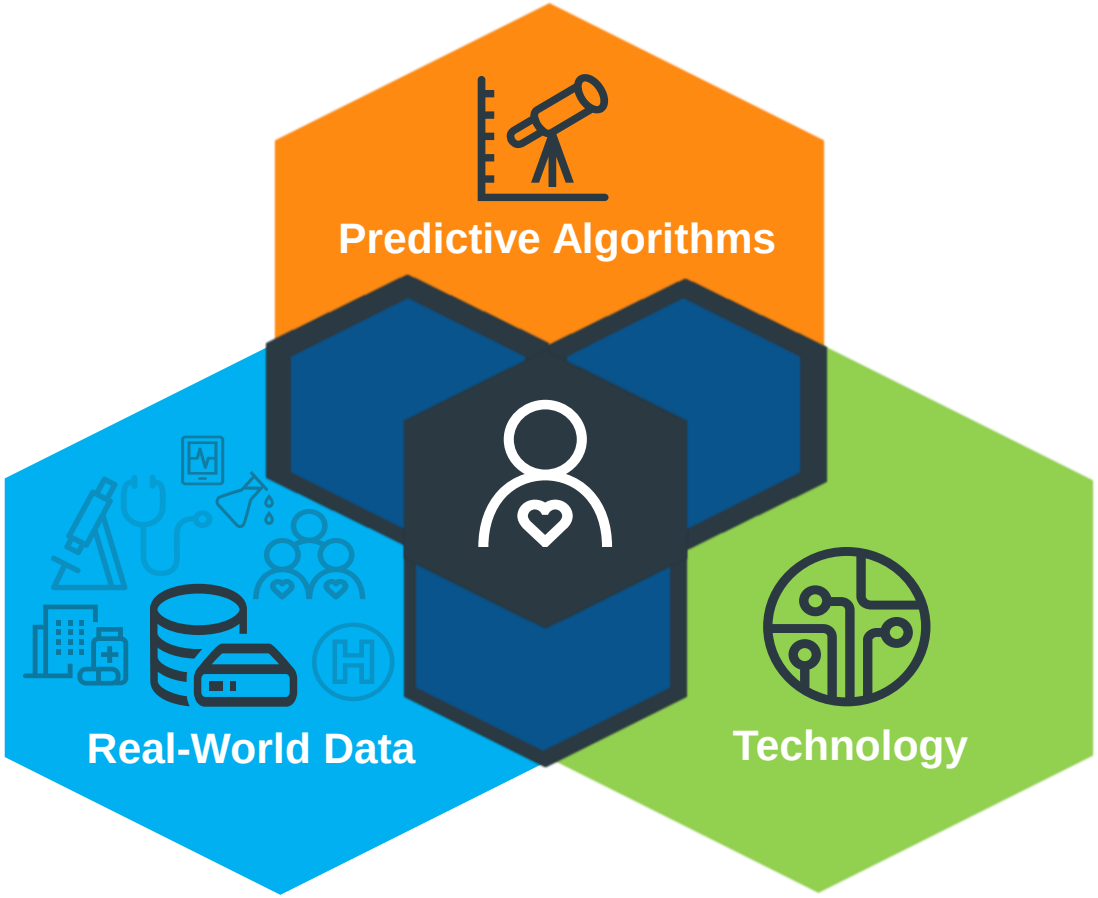
PSI 2018

Orla Doyle, PhD,
Predictive Analytics, RWAS, IQVIA

Perfect time for predictive analytics

Discovery of new analytical methods, such as pattern recognition, can exploit often complex, subtle patterns in the data to uncover new insights

Rich RWD is becoming ever more available and complex covering patient symptomatology, diagnoses, treatment history, lab tests, etc.



Advances in server infrastructure are creating new opportunities for faster processing of petabytes of data

Predictive analytics, machine learning and artificial intelligence: How do they fit together?

Artificial intelligence (AI)

"*AI involves machines that can perform tasks that are characteristic of human intelligence"* John McCarthy, 1956.

- Broad concept of using "smart" algorithms to carry out tasks.
- It used to solve tasks like object and audio recognition, learning and problem solving.
- Machine learning is a subset of AI

What is Predictive Analytics?

- Any statistical method used for predictive analysis – the practice of extracting information from data to determine patterns and predict future outcomes
- Focused on prediction at the observation-level as opposed to group-level
- Encompasses classical statistics, machine learning and Artificial Intelligence

Comparing classical statistics and machine learning

Description	Scientific philosophy	Primary objective	Example techniques	Strength	Limitation
Classical Statistics	Hypothesis driven (deductive)	Inference	Linear regression, decision tree	Well understood, transparent	Overlook complex associations
Machine Learning	Empirically-driven (inductive)	Prediction	Random Forest, Support Vector Machine	Accurate predictions in high dimensional/ complex data	Partial loss in transparency

Machine learning can transform accuracy of predictions, but only where data is complex / high dimensional

What is meant by complex data (selected examples)?

- Non-linearities: Associations between key attributes and outcome is not linear
- Non-additivity: Associations between combinations of different attributes and outcome is important
- Low degrees of freedom: High ratio of attributes to patients
- Machine learning is good at producing accurate predictions when data is complex
  - Flexible / non-parametric approach with in-built regularisation
- Dimensionality of data can often be taken as proxy for complexity
  - Dimensionality: Number of attributes and / or ratio of attributes to patients
- Thus, machine learning typically provides boost in accuracy in high dimensional data

Machine learning can increase predictive accuracy in high dimensional data

Prediction of whether a patient…	Number of attributes (apx)	Dimensionality	Predicted Positive Value (PPV) at 50% Sensitivity		Significance
			Traditional method	*Machine learning method*	
Initiates labour following induction	20	Low	77%	74%	N
Non-adheres to statins medication	50	Low/Medium	60%	64%	N
Has Hepatitis C	300	Medium	2%	87%	Y
Transitions to next line therapy in lung cancer	300	Medium	41%	89%	Y
Has a rare oncology disease	100,000	High	8%	21%	Y

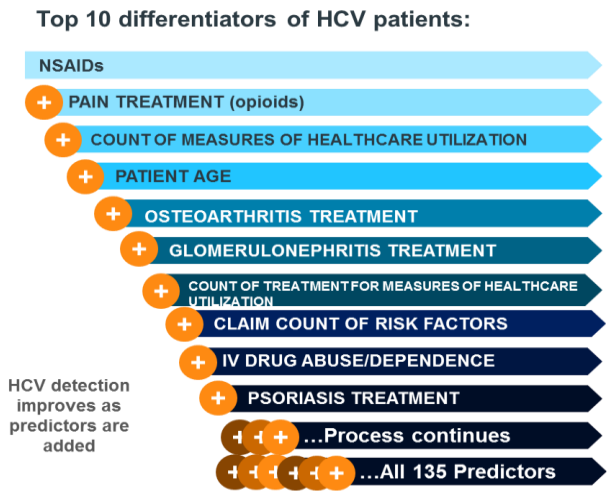
The trade off between model transparency and performance is usually a key consideration in machine learning in healthcare

Why is transparency important?

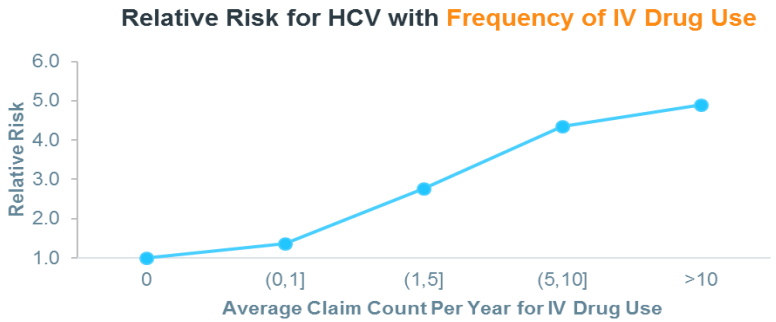
- Machine learning is empirically-driven, i.e. bottom up. Therefore interpretation is key for providing the end user with not only the “who” but also the “why” as well as for assessing of bias and engendering user trust.
- Maximising model performance can involve some loss in transparency; e.g. linear to non-parametric model, ensemble methods.
- Hence there is often a trade off. The extent to which transparency is prioritized is dependent on the application. For e.g. in speech recognition performance is priority, however, for a model identifying genomic associations with clinical outcomes transparency is a priority.

Machine learning models can sometimes be referred to as black boxes. However, several model agnostic approaches are readily available.

Feature importance: how does model performance change when a feature is removed?



Relative risk: how does the prediction change as a feature value changes?

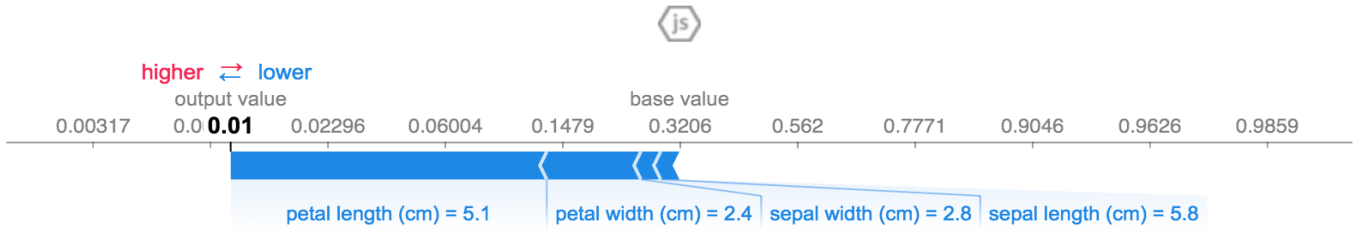


Patient profiles: what are the key attributes of the top high risk patients?

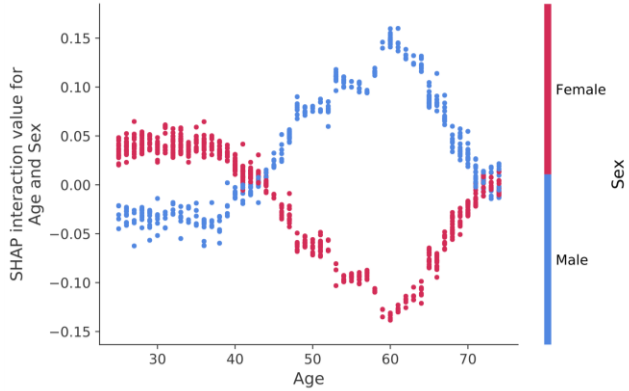
PREDICTORS AVERAGE CLAIM COUNT	John Doe 1 Age: 69 Gender: M	John Doe 2 Age: 50 Gender: M	Jane Doe 1 Age: 25 Gender: F	Jane Doe 2 Age: 49 Gender: F
NSAIDS	●	●	●	●
PAIN TREATMENT	●	●	●	●
OSTEOARTHRITIS TREATMENT	●		●	●
GLOMERULONEPHRITIS TREATMENT	●		●	
RISK FACTORS			●	
IV DRUG USE			●	
RHEUMATOID ARTHRITIS	●	●	●	●
THROMBOSIS TREATMENT PROC	●			
DEPRESSION TREATMENT			●	
NAUSEA TREATMENT	●		●	
ANXIETY TREATMENT	●		●	●
HIV/AIDS				
ALCOHOL ABUSE				
CLAIM COUNT OF COMORBIDITIES	30	0	11	5
OVERALL COUNT OF PREDICTORS	27	9	21	18

The machine learning field is transforming black box models to glass models with transparency as an active area of research

SHAP (SHapley Additive exPlanations) is a unified approach to explain the output of any machine learning model. It provides a wide range of insights including:



1 The contribution of features contribute to a single prediction.



2 The interaction between variables for predictions.

The **LIME** algorithm approximates a black-box model by a simple local model (i.e. in the neighborhood of the prediction we want to explain), as opposed to trying to approximate a model globally (i.e. on all samples).

1 The training images contain cats and dogs.

2 The test image contains a cat and a dog.

3 Examples of cats and dogs have been seen in training. LIME can localise the pixels related to the “dog” class and separately those related to the “cat” class.

SHAP images sourced from https://github.com/slundberg/shap
LIME images sourced from https://marcotcr.github.io/lime/tutorials/Tutorial%20-%20images.html

We have deep experience across therapeutic categories and using many types of data sources in both US and Ex-US markets

AI and ML library of best practices and expertise on largest curated medical data assets in the world – ability to also incorporate EMR, patient registries, consumer data, etc.

Screening / Disease Detection

Hepatology

Hepatitis C
NASH

Miscellaneous

Exocrine pancreatic insufficiency
Endometriosis
Pulmonary arterial hypertension

Respiratory

Idiopathic pulmonary fibrosis
Nontuberculosis mycobacterium
Severe asthma

Genetic

Genetic hypophosphatemia
Hereditary angioedema
Rare kidney disease

Neurological

Tardive dyskinesia
Primary periodic paralysis

Predicting response / line of therapy transition

Oncology

Line of therapy change in:
Small cell lung cancer
Prostate cancer

Miscellaneous

Line of therapy change in:
Rheumatoid arthritis
Multiple sclerosis

Ophthalmology

Treatment response in age-related macular degeneration

Obstetrics

Treatment response in induction of labour

Adherence

Predicting adherence in:
Asthma biologic market
Hyperlipidemia market

The following slides will focus on a example of disease detection in Hepatitis C.

Hepatitis C (HCV) is a **serious illness** and is considerably **under-diagnosed**

HCV is a viral infection of the liver that is acquired through contact with contaminated blood that contains the virus (i.e., blood transfusion, IV drug abuse, sexual contact)

HCV IN THE UNITED STATES

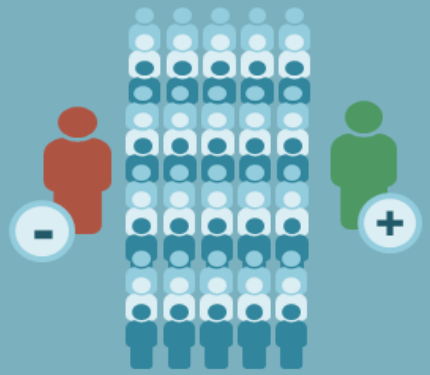
Approximately

3.5 million

people are infected

8,000-10,000

deaths each year



30,000

new cases each year

THE SILENT DISEASE

40-85%

Of persons infected are unaware of their HCV infection status

For each new case of acute symptomatic HCV

3.3

cases of acute symptomatic HCV occur

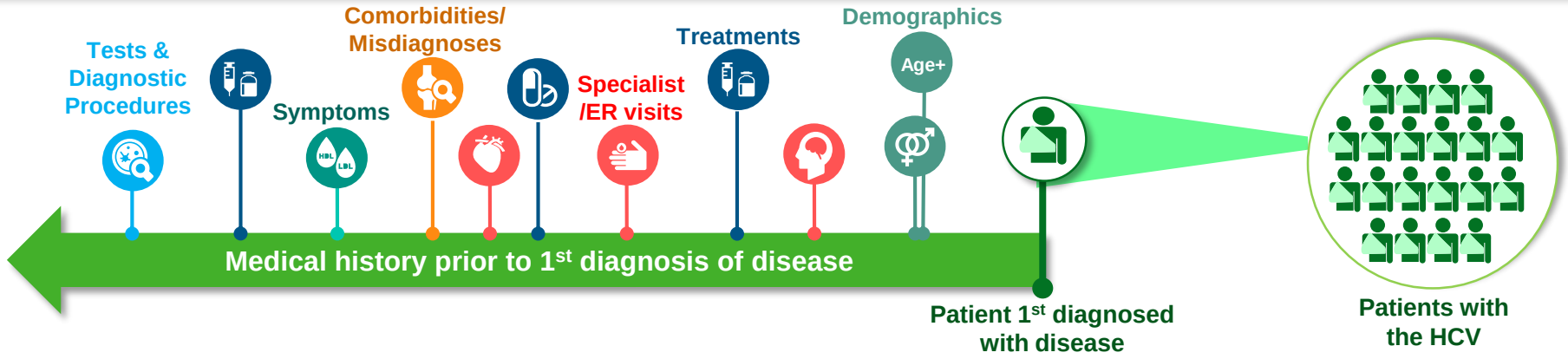
For each new case of acute symptomatic HCV that is reported

13x

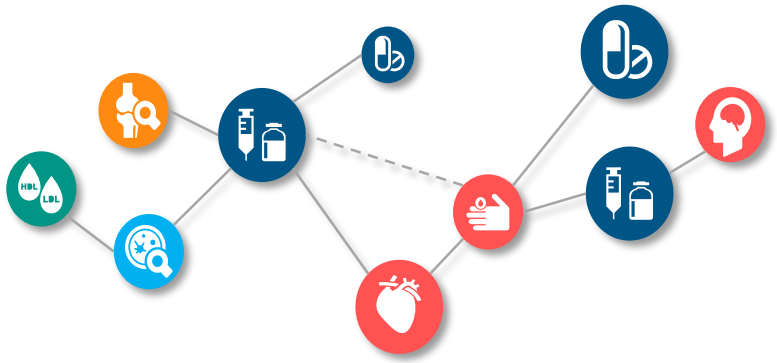
of acute symptomatic HCV that occur

Using the digital footprint of *diagnosed* patients to find *undiagnosed* HCV patients

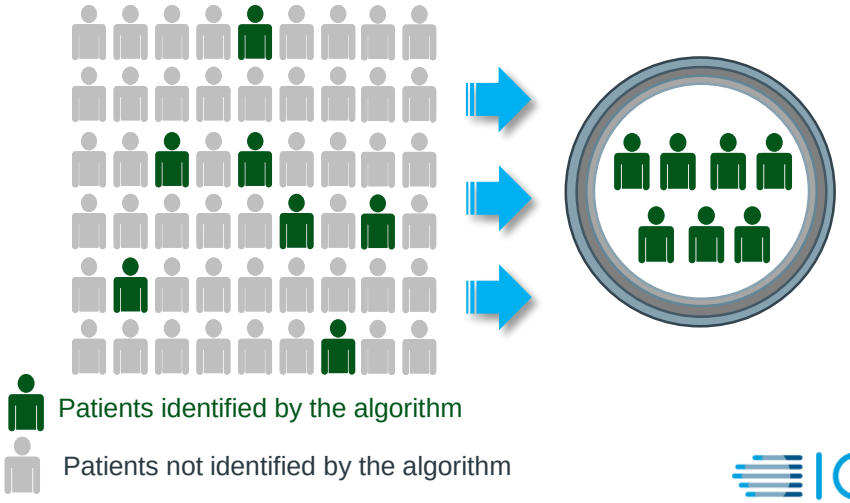
1 Identify patients with the disease and analyze their medical history PRIOR to the 1st HCV-related event in their history



2 Develop an algorithm to identify unique patterns of HCV in patients' pre-diagnosis medical history



3 Find and target patients in the wider universe who are identified by the algorithm as potentially undiagnosed

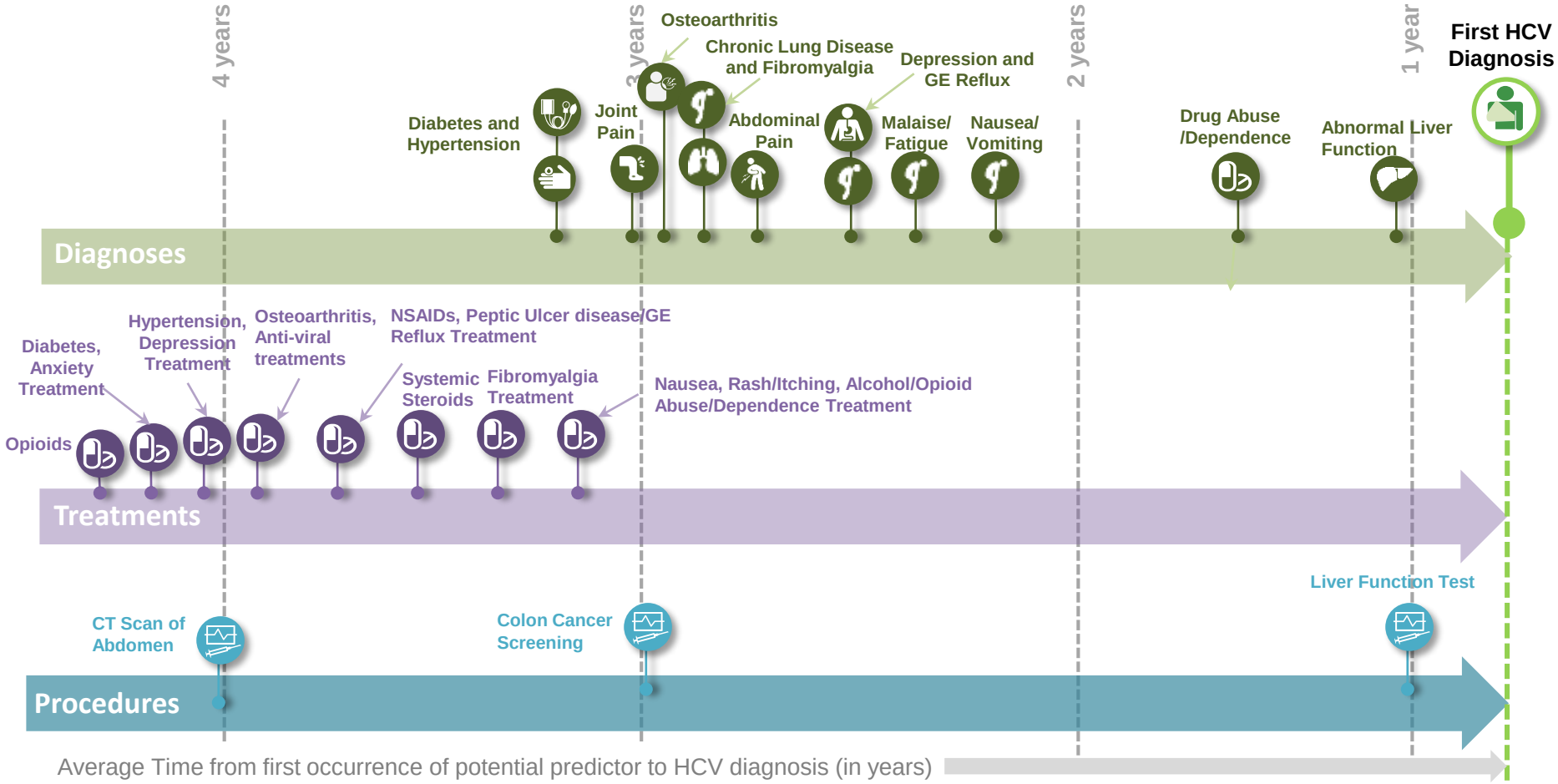


1

Timeline of HCV Patient Journey Prior to Diagnosis

Pre-diagnosis view of HCV patient journey timeline provides

- A comprehensive view of the various touch-points that a patient has with the medical system
- Valuable information on when patients start experiencing medical events and for how long they experience these events

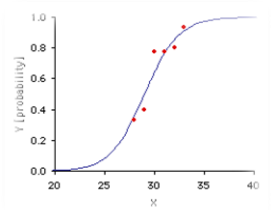
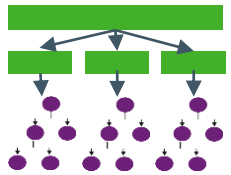
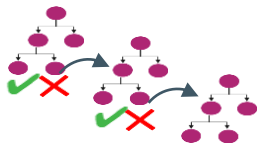
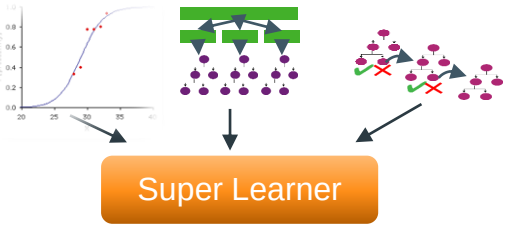


Key Findings:

- Patients begin experiencing known symptoms of HCV (joint pain, abdominal pain, malaise/fatigue, fibromyalgia) 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 trying to treat for these symptoms earlier
- Patients are undergoing several diagnostic test procedures close to the time of their diagnoses

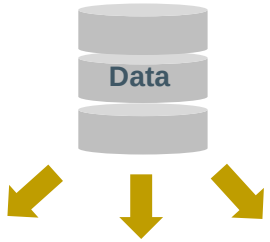
Variable lookback window and nature of open-source claims should be kept in mind when interpreting results

2 Develop an algorithm to identify unique patterns of the disease in patients' pre-diagnosis medical history

Model		Overview
Logistic regression		• Logistic regression is a well-known binary classification model • The output of a logistic regression is very interpretable as it assigns an odds ratio to each predictor (independent variable)
Random Forest		• The random forest method combines the output from hundreds or thousands of decision trees (a forest of trees) • Each tree is grown on a random sample of data with replacement • Each split is made on a random sample of input variables
Gradient Boosting Trees (xgboost)		• Similar to Random Forest, Boosted Trees combines output from many trees • Each tree is trained on a weighted sample of the data, where higher weights are assigned to observations misclassified by the previous tree • The process is particularly adept at classifying 'hard-to-classify' observations
Super Learner		• An ensemble classifier combines output from several individual classifiers. This can be achieved using a 'super-learner' which uses the predictions from subsequent models as its features. • A super-learner has the potential to capture different properties from different methods to form a single classifier which can improve overall predictive.

2 Patient data split into training sample to develop the model, and testing sample to assess the model

Model Building Approach – Data Samples for Training, Validation and Hold Out.
Training data are used to develop the model and to perform grid search for any model hyperparameters.
Validation data are used for model selection.
Hold Out data are used to assess the performance of the chosen model on independent data.



Training Sample (80%)
To develop and train the model

HCV Patients *Non-HCV Patients*

Positive Cohort Negative Cohort

Training used a HCV to Non-HCV ratio of 1:50. That is the majority class were under-sampled in training to help alleviate the challenge with class imbalance.

Validation Sample (10%)
To evaluate model performance for model selection

HCV Patients *Non-HCV Patients*

Positive Cohort Negative Cohort

Testing used an HCV to Non-HCV ratio of 1:200*. This ensures that the model is validated on data which reflect the likely prevalence.

Hold Out Sample (10%)
To evaluate performance of the chosen model

HCV Patients *Non-HCV Patients*

Positive Cohort Negative Cohort

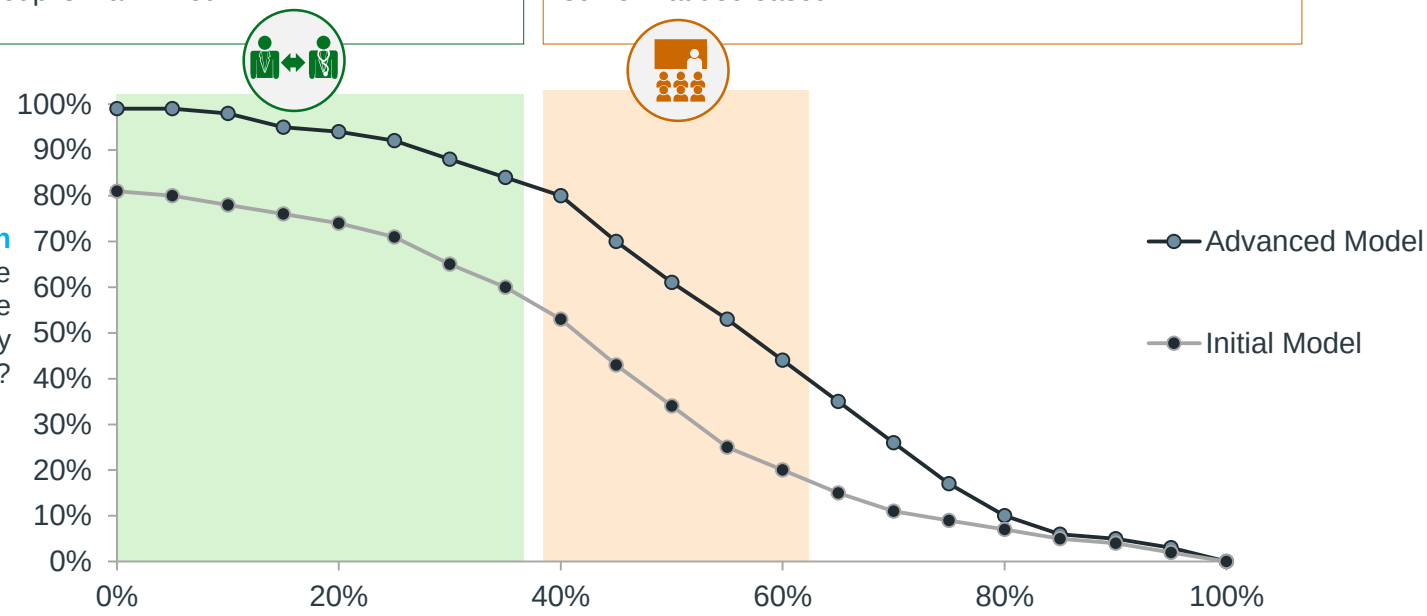
Testing used an HCV to Non-HCV ratio of 1:200*. This ensures that the model is validated on data which reflect the likely prevalence.

**Representing a conservative estimate of the prevalence of undiagnosed HCV in the general population (diagnosed prevalence is ~3.5 million in the US).*
Cohorts were further stratified prior to modeling to improve model precision

2 Model development is typically an iterative process involving the comparison of different methods

Highly Confident Predictions
A modest number of patients are identified as HCV however the likelihood of belonging to the HCV group is maximized.

Moderately Confident Targets
An increased number of patients are identified as HCV however the likelihood of belonging to the HCV group is somewhat decreased.



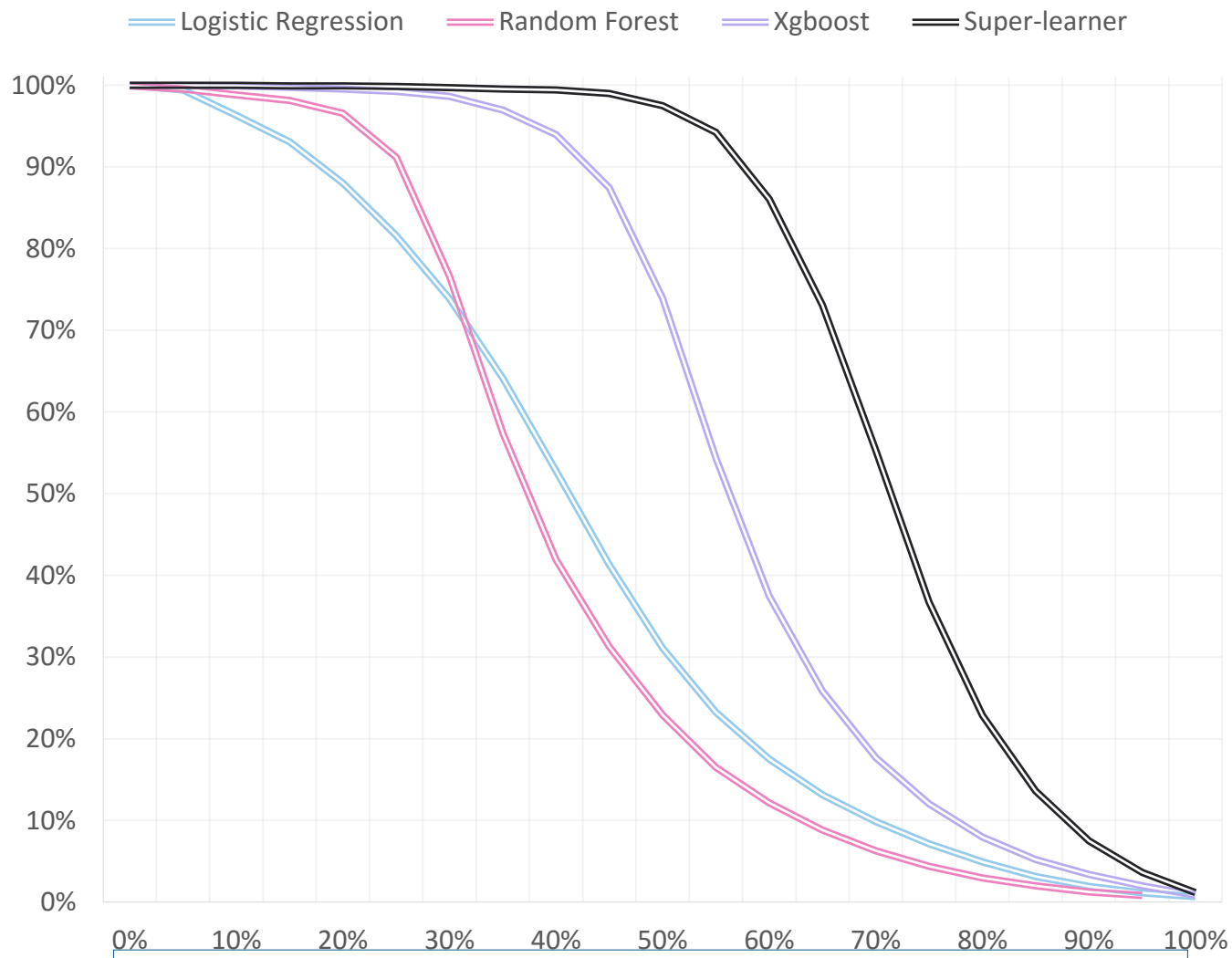
Precision
Of all patients who were flagged, what % were actually subsequently diagnosed with HCV?

Recall
What % of HCV patients were flagged by the model?

- For any model, there is a trade-off between precision and recall
- Selecting the optimal balance will in part depend on client's goals
- Better models shift the curve up to the right, i.e. maximizing recall for a given level of precision
- IQVIA will identify the model with highest precision for a given recall

2

Machine learning methods were found to be highly effective at finding undiagnosed HCV patients



Precision
Of all patients who were flagged, what % went to receive a diagnosis of HCV?

Recall: What % of actual HCV patients were flagged by the model?

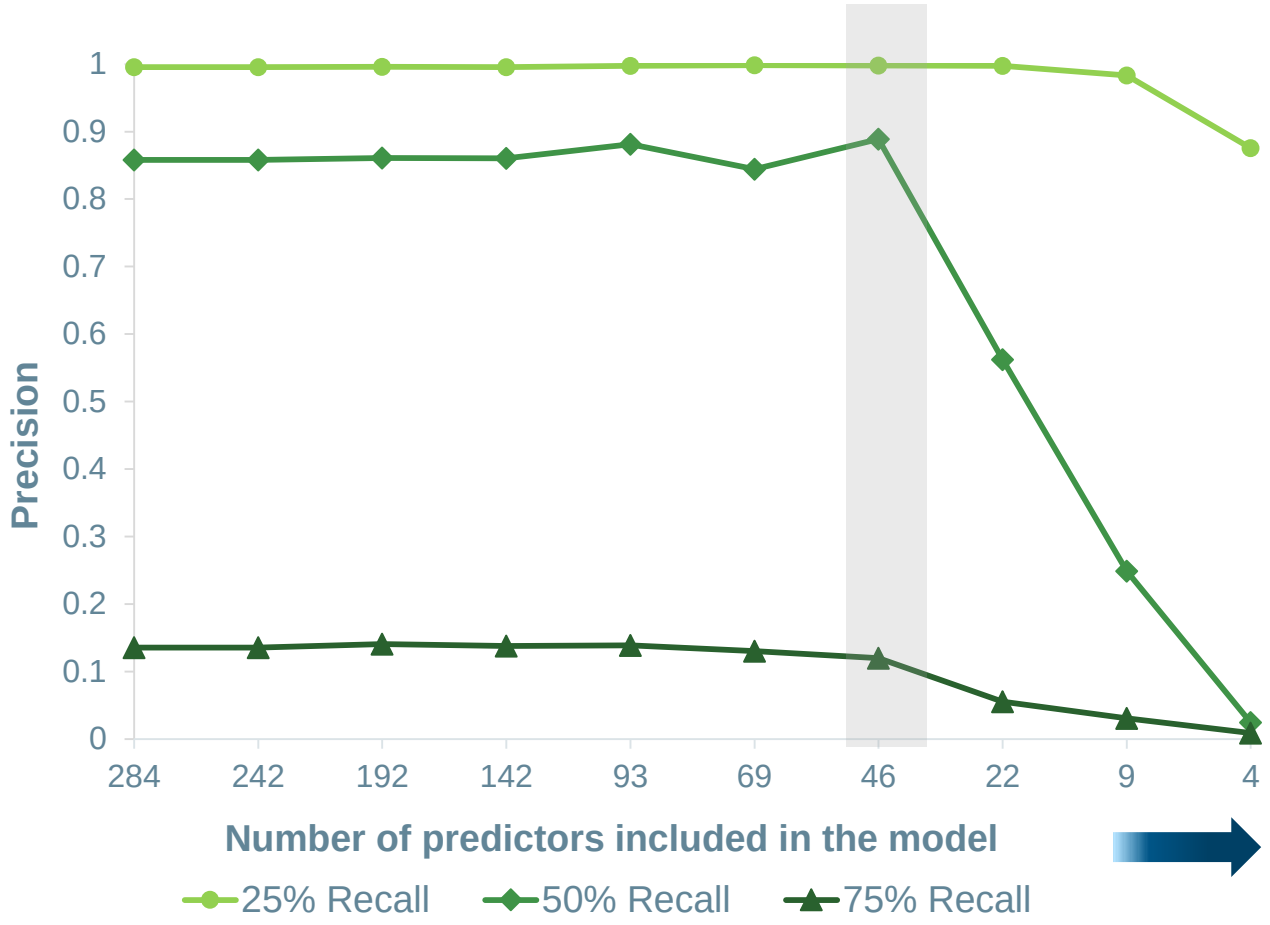
The super-learner achieved superior performance to all algorithms and the PR-curve presented was that achieved on the hold-out sample.

Key findings:

- At low levels of recall (<20%) the difference in performance across algorithms in modest.
- At higher levels of recall (50%) the difference is much more profound with the super-learner achieving a precision of 97% compared to 74% for the next best performing model.
- That is, for 50% recall, 97 out of 100 patients flagged by the model were subsequently diagnosed with HCV shortly afterwards.

2

Substantially reducing the number of predictors does not affect precision



Model with 46 predictors:

99% Precision at 25% Recall

*If the goal is to find 25% of HCV patients then **almost all patients** predicted to be at risk of having HCV are true positives*

89% Precision at 50% Recall

*If the goal is to find 50% of HCV patients then approximately **9 out of 10** patients predicted to be at risk of having HCV are true positives*

12% Precision at 75% Recall

*If the goal is to find 75% of HCV patients then approximately **1 out of 10** patients predicted to be at risk of having HCV are true positives*

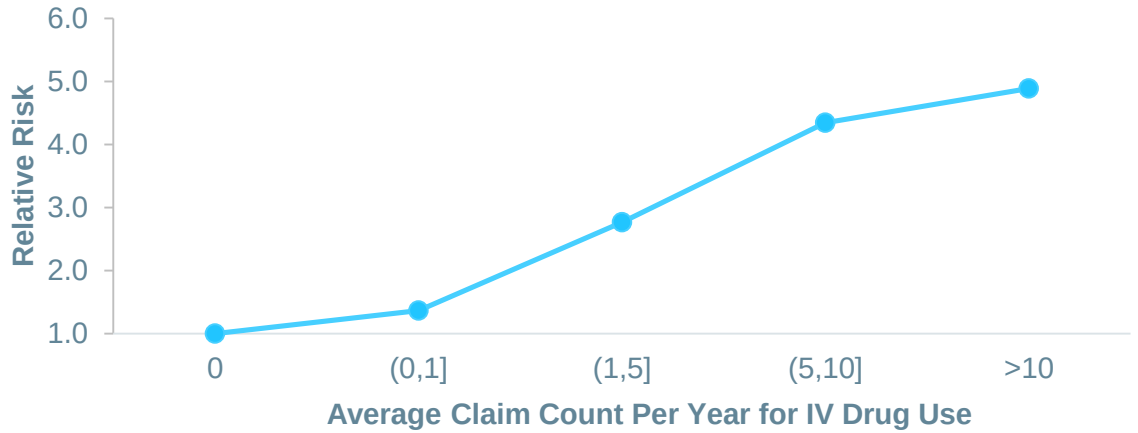
2 Machine learning models also reveal the key attributes of an undiagnosed HCV patient

Top 10 differentiators of HCV patients:

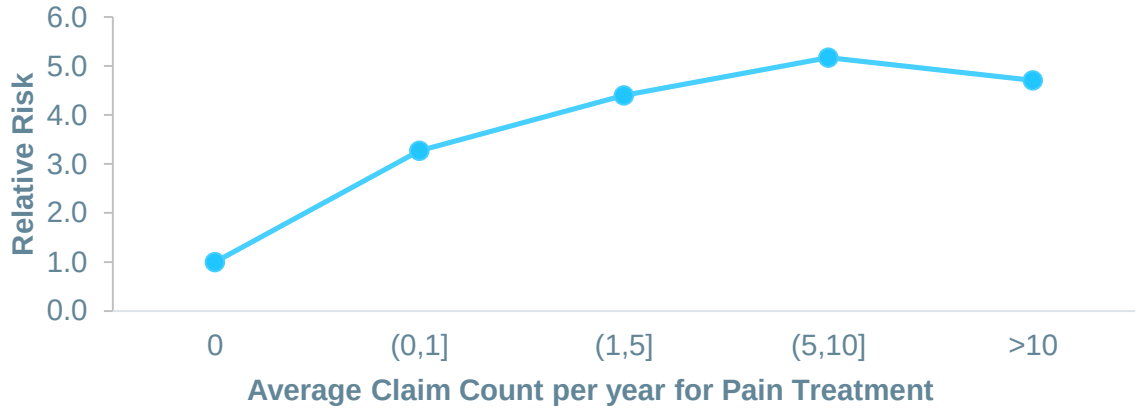
- NSAIDs
- + PAIN TREATMENT (opioids)
- + COUNT OF MEASURES OF HEALTHCARE UTILIZATION
- + PATIENT AGE
- + OSTEOARTHRITIS TREATMENT
- + GLOMERULONEPHRITIS TREATMENT
- + COUNT OF TREATMENT FOR MEASURES OF HEALTHCARE UTILIZATION
- + CLAIM COUNT OF RISK FACTORS
- + IV DRUG ABUSE/DEPENDENCE
- + PSORIASIS TREATMENT
- + + + …Process continues
- + + + + + + …All 135 Predictors

HCV detection improves as predictors are added

Relative Risk for HCV with Frequency of IV Drug Use



Relative Risk for HCV with Frequency of Pain Treatment



IQVIA™

2 Profiles of highest-risk patients for HCV show several emerging patterns

PREDICTORS AVERAGE CLAIM COUNT	John Doe 1 Age: 69 Gender: M	John Doe 2 Age: 50 Gender: M	Jane Doe 1 Age: 25 Gender: F	Jane Doe 2 Age: 49 Gender: F
NSAIDS	●	●	●	●
PAIN TREATMENT	●	●	●	●
OSTEOARTHRITIS TREATMENT	●		●	●
GLOMERULONEPHRITIS TREATMENT	●		●	
RISK FACTORS			●	
IV DRUG USE			●	
RHEUMATOID ARTHRITIS	●	●	●	●
THROMBOSIS TREATMENT PROC	●			
DEPRESSION TREATMENT			●	
NAUSEA TREATMENT	●		●	
ANXIETY TREATMENT	●		●	●
HIV/AIDS				
ALCOHOL ABUSE				
CLAIM COUNT OF COMORBIDITIES	30	0	11	5
OVERALL COUNT OF PREDICTORS	27	9	21	18

FINDINGS

John Doe 1:
High amount of comorbidities and pain treatment

John Doe 2:
Pain treatment only

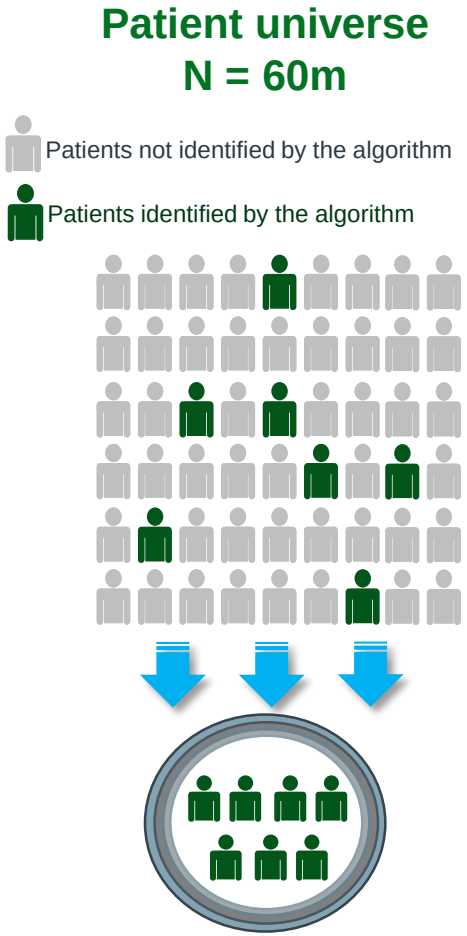
Jane Doe 1:
Basically all risks, including age

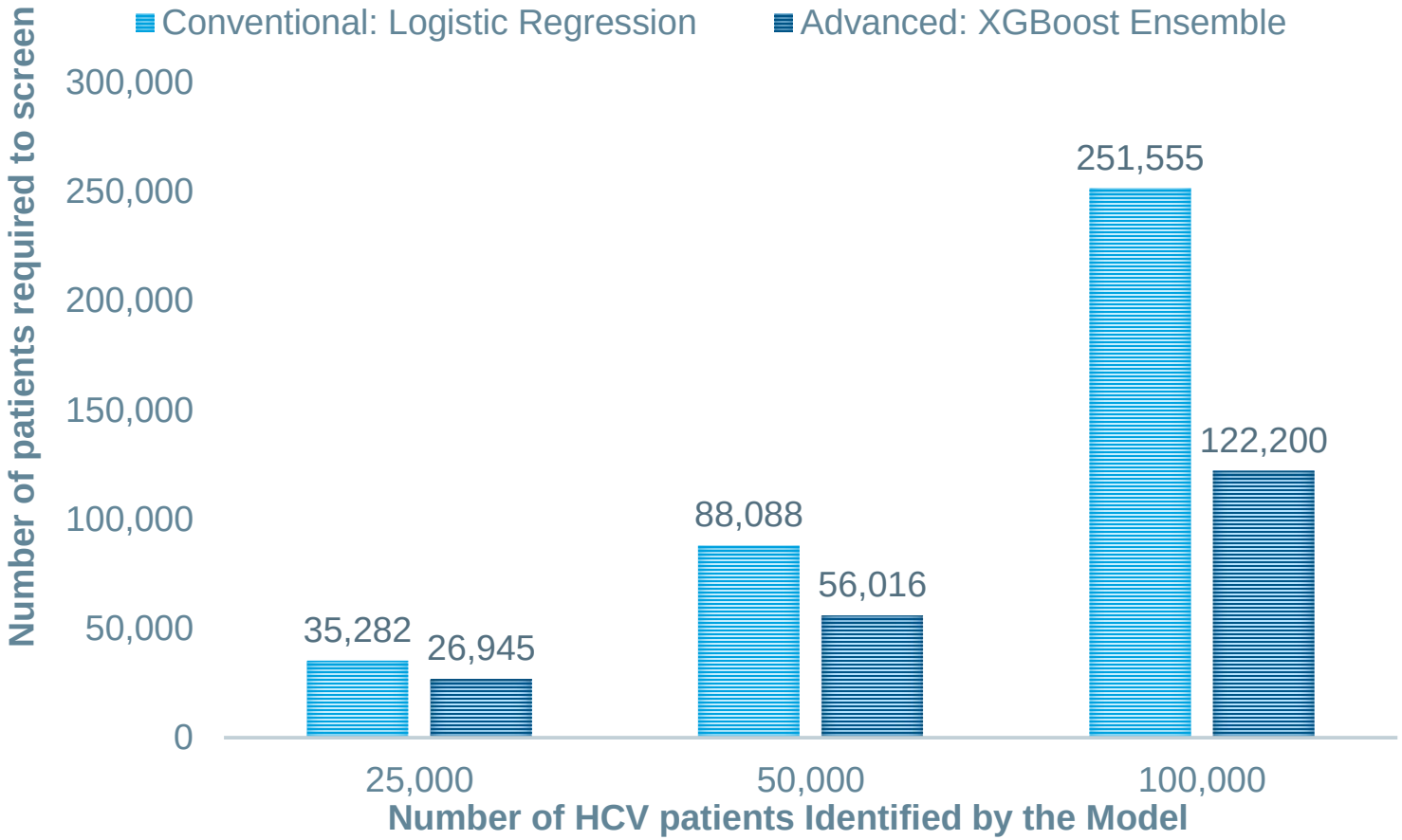
Jane Doe 2:
Pain treatment and psychiatric illness

Note that no highest-risk patient had HIV/AIDS or alcohol abuse.

3 Apply the algorithm to a completely independent patient universe to identify undiagnosed patients

The advanced machine learning approach we developed requires 122k patients to be screened in order to find 100k undiagnosed HCV patients – more than 2x better than a conventional logistic regression model





Challenges and opportunities for payers and providers relating to patient-finding applications (selected examples)

Challenge	Opportunity
Uncertain value proposition e.g. which intervention is most effective for which patient?	Clear use cases with value propositions required – which patients benefit most and why
Decentralised and fragmented health systems leading to patchy coverage of patient journeys and high cost for widespread roll-out of applications	Increased linkage within and between different health systems
Lack of detailed biological data (esp. omics) constraining precision medicine applications	Routine biological sampling with results integrated into EHR
Regulatory uncertainty – especially for precision medicine – How should algorithms be validated and updated in a timely fashion?	Regulatory clarity and practical use cases required

IMS Health & Quintiles are now IQVIA™

Thank you!

For further information, please contact:
Orla Doyle, PhD
Senior Data Scientist,
Predictive Analytics,
Real-World & Analytics Solutions
IQVIA
Orla.Doyle@IQVIA.com