SAT-331

Use of Machine Learning to Predict Diagnosis Codes for Nonalcoholic Steatohepatitis in Administrative Healthcare Data

Gilead Sciences, Inc.
333 Lakeside Drive
Foster City, CA 94404
1-800-445-3235

Laura E. Telep[1], Chi Zhang[2], Stephen Djedjos[1], Robert P. Myers[1], Robertino Mera[1], Alan Hubbard[2], Anand P. Chokkalingam[1,2]

[1]Gilead Sciences, Foster City, CA USA; [2]School of Public Health, University of California, Berkeley, CA, USA

Background and Aims

- Background
  - The natural history of nonalcoholic steatohepatitis (NASH) is poorly understood
  - Analysis of US administrative claims data to characterize the long-term consequences of NASH has been hampered by the grouping of NASH with other nonalcoholic fatty liver disease (NAFLD) in version 9 of the International Classification of Disease (ICD-9).
  - In ICD-10 (adopted in the US in 01 Oct 2015), NASH has a unique diagnostic code
- Aims
  - To use machine learning to identify ICD-9 NAFLD/NASH patients likely to have a claim for ICD-10 NASH based on claims observed in the ICD-9 era
  - To create a cohort of NASH patients which spans ICD-9 and ICD10 claims as the expanded follow up time will enable a better understanding of the natural history of this condition

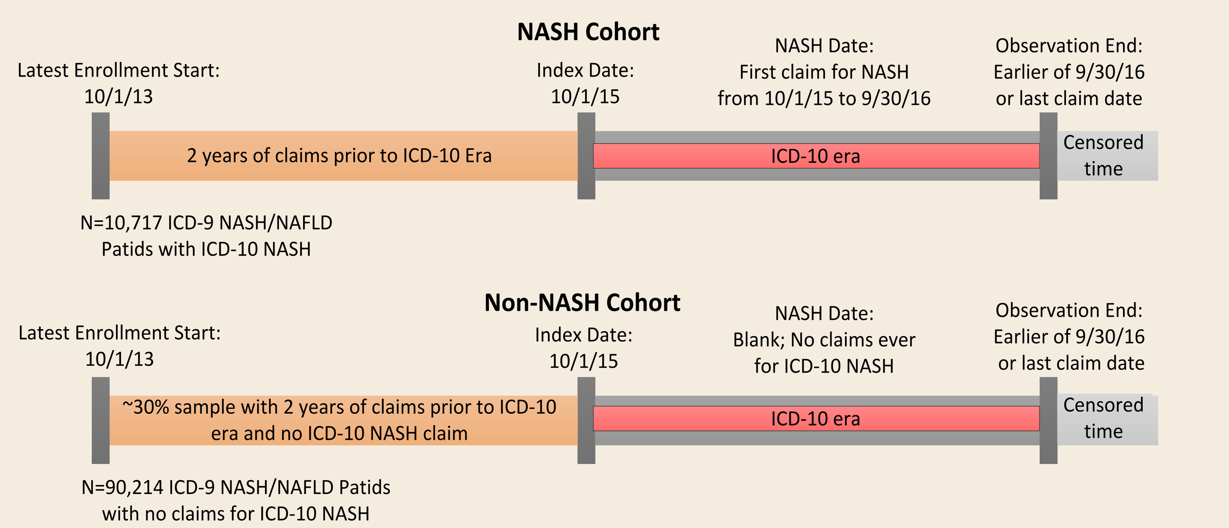
Methods

- Approach
  - Use the ensemble method Super Learner (SL) with leave one group out cross validation (LOGO CV) to create an algorithm which identifies ICD-9 NAFLD/NASH patients who would be likely to have a claim for ICD-10 NASH
  - Apply this algorithm to cohorts of patients with claims in the ICD-9 era
- Data source:
  - IQVIA PharMetrics Plus™ Claims dataset, including U.S. administrative claims for ~140M patient lives from 01 Jan 2006 to 30 Mar 2018
    - Claims from 01 Jan 2006 – 30 Sep 2015 reported with ICD-9 codes
    - Claims form 01 Oct 2015 – 30 Mar 2018 reported with ICD-10 codes

Data Set for Predictive Algorithm Creation

- Every patient included in the data set:
  - had an ICD-9 claim for NAFLD/NASH (571.8)
  - was at least 18 years of age on 01 Oct 2015 (index date)
  - had continuous enrollment during the final 2 years of the ICD-9 era (01 Oct 2013 – 30 Sep 2015)
- NASH cohort patients (N=10,717) had a claim for NASH (K75.81) in the first year of the ICD-10 era (01 Oct 2015 – 30 Sep 2016)
- The Non-NASH cohort (N=90,214) was comprised of a random ~30% sample of patients with no claim for ICD-10 NASH in any observation time in the ICD-10 era (starting 01 Oct 2015)

Data Set for Predictive Algorithm Creation



Claim Code Selection

- Identify all distinct ICD-9 diagnosis codes and GPI medication codes occurring during the 2 years of claims prior to the ICD-10 era (01 Oct 13 – 30 Sep 15)
  - Create grouped variables to represent 10 pre-specified conditions known to be associated with NASH from
    - 133 distinct diagnostic codes
    - 78 distinct medication codes
  - Any claim code not included in a pre-specified conditions is considered individually
- Create a 1/0 flag for each grouped or individual claim code that occurs during the 2 year observation prior to the ICD-10 era
- Final list of variables considered for NASH prediction:
  - Age
  - Sex
  - 10 pre-specified covariates
  - 2,683 diagnosis codes
  - 470 medication codes

Methods (cont'd)

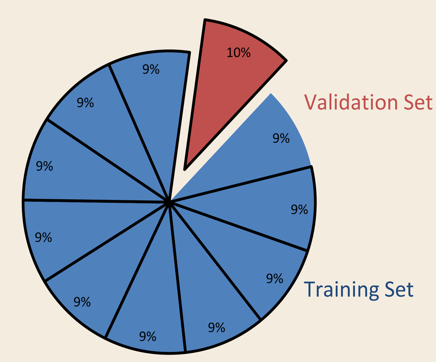
10 Pre-Specified Covariates

Pre-Specified NASH Covariates		
Condition	**ICD-9 COdes**	**GPI Codes**
Obesity	278.0x	
Diabetes	C	27x
Metabolic Syndrome	277.7x	
Hyperlipidemia	272.0x, 272.2x, 272.4x	39x
Hypertriglyceridemia	272.1x	
Stroke	434.x	
Essential Hypertension	410.x	
Myocardial Infarction	410.x, 412.x	
Coronary Atherosclerosis	414.0x, 414.3x, 414.4x, 429.2x	
Smoking	305.1x, V15.82x	

Note: here "x" refers to any subsequent digits in the ICD-9 or GPI ontology

Super Learner Ensemble Method Setup

Full data set is divided: Training Set (90%) and Validation Test set (10%)



Step 1: Discrete Super Learner run on 90% of Training Set, validated using LOGO CV (10-Fold)

- Screen claims codes :
  - Utilize Bayesian Risk Ratio (BRR) for each code and select inclusion thresholds
    - BRR > 1.47 OR BRR < 0.68
  - Implement Sparsity Thresholds
  - Run LASSO regression (Least Absolute Shrinkage and Selection Operator) to help with variable selection and regularization
- Run each of 19 selected prediction algorithms (base learners) individually to identify those that contribute to the final model

Step 2: Meta-Learner run on full Training Set

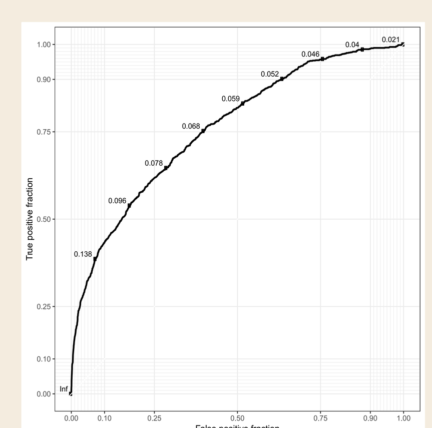
- Repeat BRR calculation, sparsity measures, and LASSO
- Run Super Learner
  - using only those base learners found to be significant when run individually during Step 1
  - Final model is a weighted combination of results from individual base learners

19 base learners for the machine learning library in Super Learner

19 Base Learners with Super Learner-Assigned Weights			
Base Learner	**Weight**	**Base Learner**	**Weight**
Generalized Linear Model (GLM)	0.077	Flexible Discriminant Analysis	0.038
Bayesian GLM	0.076	Classification and Regression Trees (CART)	0.000
Elastic-Net Regularized GLM	0.077	Multivar. Adaptive Regression Splines (Param Set 1)	0.011
Boosted GLM	0.021	Multivar. Adaptive Regression Splines (Param Set 2)	0.015
Penalized Multinomial Regression	0.080	Penalized Discriminant Analysis (Param Set 1)	0.066
Boosted Logistic Regression	0.041	Penalized Discriminant Analysis (Param Set 2)	0.035
Naïve Bayes	0.056	Single C5.0 Rule-Based Models	0.058
Nearest Shrunken Centroids	0.027	Single C5.0 Decision Tree Models	0.119
Shrinkage Discriminant Analysis	0.066	Conditional Inference Tree	0.000
Stochastic Gradient Boosting	0.136		

*Tuning parameters were selected with 5-fold cross validation

- Final Predictive Algorithm Based On:
  - 17 of the 19 base learners (those with non-zero weights)
  - 371 ICD and GPI codes, including
    - 211 pre-specified, based on 10 combined codes
    - 160 agnostically selected codes, based on BRR
  - Age
  - Sex
- Final AUC = 0.76



Data Sets to which the Predictive Algorithm was Applied

- Cohorts of patients were developed in a manner consistent with the development of the predictor:
  - Minimum of 2 years of claims and an ICD-9 claim for NASH/NAFLD in the period from 01 Oct 2010 to 31 Aug 2015
    - versions of ICD-9 claim codes prior to 01 Oct 2010 did not include some codes which were identified by the predictor as indicative of NASH
  - The prediction algorithm produced a percent likelihood of NASH at the end of 2 years of observation based on condition and medication claims during that time period
  - Patients were required to have a minimum of 30 days for post-prediction observation in the ICD-9 era
- Prediction thresholds selected based on algorithm performance in Validation Test Set:
  - Youden's index (sensitivity+specificity-1)
  - Specific PPV thresholds: 20%, 35%
- Once a patient was predicted to have NASH, this patient was removed from future cohorts for NASH prediction
- If a patient was predicted to not have NASH in the first 2-year window of available claims data, and an additional year of claims (plus at least 30 days) was available, the observation window was advanced forward one year and the predictor was re-applied

Results

Prediction Cohort sizes at different PPV Thresholds

Prediction Threshold	Number of patients
Total eligible	**405,233**
Youden Index (8% PPV)	216,180
20% PPV	41,912
35% PPV	10,771

Selected Outcomes: Predicted NASH Cohort vs. Claimed ICD-10 NASH Cohort

Methods:

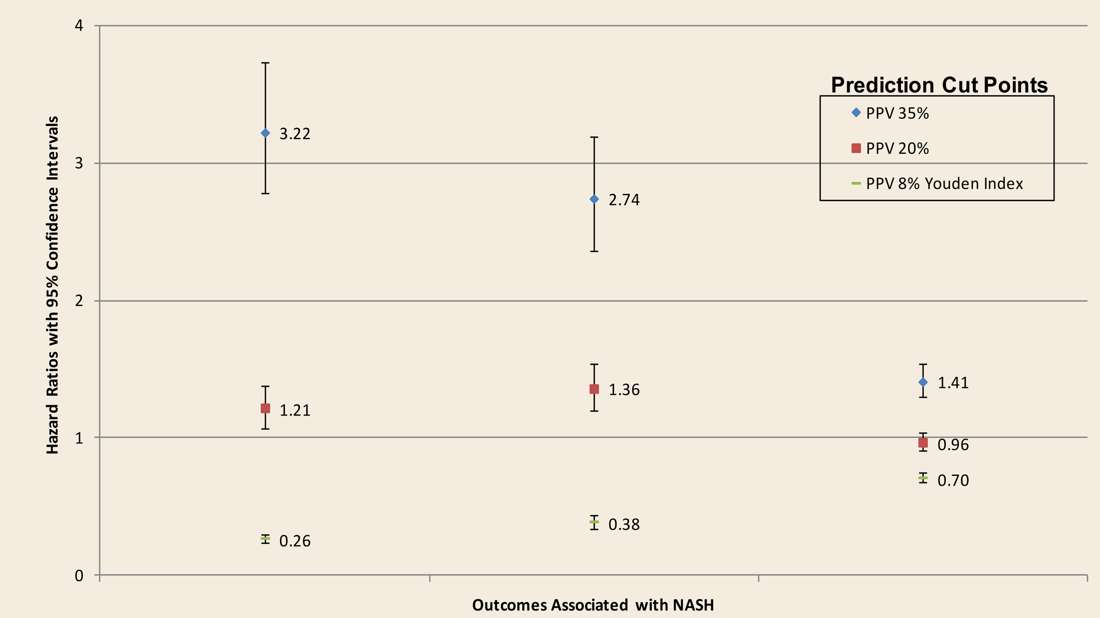
- Data source: IQVIA PharMetrics Plus™ Claims dataset, including US administrative claims for ~140M patient-lives from 01 Jan 2006 to 30 Mar 2018
- Index date: predicted NASH date (in predicted cohort) or first ICD-10 NASH claim (in claimed ICD-10 NASH cohort)
- Age ≥18 years on index date with no prior claims for the outcome of interest
- End of follow-up for all patients was defined as the earliest of:
  - Date of first claim for outcome of interest
  - Date of last claim in data set
  - End of data (30 Sep 2015 for predicted cohort, 30 Mar 2018 for ICD-10 NASH cohort)
- Hazard ratios (HRs) and corresponding 95% confidence intervals (CIs) calculated using Cox proportional hazards methods
  - Time-to-event analysis

Selected Outcomes: Predicted NASH Cohort vs. Claimed ICD-10 NASH Cohort

Variable Definitions	Codes		
Outcome:	**ICD-9-CM**	**ICD-10-CM**	**CPT/HCPCS**
HCC	155.0x, 155.2x	C22.0x, C22.2x, C22.3x, C22.4x, C22.7x, C22.8x, C22.9x	
Cardiac Events (Acute and Chronic ischemic heart disease, Angina pectoris, Cardiomyopathy, Heart Failure)	410.X (except with a 5th digit of 2), 411.x, 413.x (except 413.1x), 414.x (except 414.1x), 425.x, 428.x	I20.x (except I20.1x), I21.x, I22.x, I24.x, I25.x (except I25.2x, I25.3x, I25.4x), I42.x, I50.x	
Liver Transplant	V42.7x	z94.4	47135, 47136, 505, 505.1, 505.9, 0093, 0FY00Z0

Results (cont'd)

Selected NASH Outcomes



	Liver Transplant			HCC			Cardiac Events		
	subjects	evts	ptime (yrs)	subjects	evts	ptime (yrs)	subjects	evts	ptime (yrs)
Predicted: PPV: 35%	9,990	385	14,613	9,986	352	14,715	6,955	741	9,943
Predicted: PPV: 20%	40,492	588	62,580	40,282	655	62,405	30,607	2,122	46,281
Predicted: PPV: 8%	214,520	650	347,326	213,976	922	346,490	170,028	8,614	267,745
ICD-10 NASH Cohort	42,376	379	48,423	42,134	365	48,258	33,648	1,748	37,434

Demographic Characteristics of Predicted NASH Cohort (PPV 20%) and Claimed ICD-10 NASH Cohort			
Variable		**Predicted NASH (N=41,912) N(%)**	**ICD-10 NASH (N=42,744) N(%)**
Age Group	18-34 y	2,263 (5.40)	3,655 (8.55)
	35-44 y	5,172 (12.34)	6,887 (16.11)
	45-54 y	11,974 (28.57)	12,262 (28.69)
	55-64 y	17,647 (42.10)	15,747 (36.84)
	65-74 y	4,189 (9.99)	3,713 (8.69)
	75+ y	667 (1.59)	480 (1.12)
Sex	Female	21,409 (51.08)	22,260 (52.08)
	Male	20,503 (48.92)	20,484 (47.92)
FU Time (years)	Mean (SD)	1.56 (0.94)	1.15 (0.71)
	Median (Q1, Q3)	1.33 (0.74, 2.23)	1.06 (0.52, 1.75)
Baseline Conditions	Obesity	16,947 (40.43)	19,369 (45.31)
	Diabetes	22,556 (53.82)	19,128 (44.75)
	Metabolic Syndrome	3,024 (7.22)	2,217 (5.19)
	Hyperlipidemia	28,781 (68.67)	28,101 (65.74)
	Hypertriglyceridemia	2,480 (5.92)	2,589 (6.06)
	Essential Hypertension	33,092 (78.96)	31,547 (73.81)
	Stroke	942 (2.25)	648 (1.52)
	Myocardial Infarction	1,787 (4.26)	1,334 (3.12)
	Coronary Atherosclerosis	5,981 (14.27)	4,400 (10.29)
	Smoking	8,702 (20.76)	8,488 (19.86)

Future Directions

Further development of NASH predictor

- Additional dimensions for prediction such as the inclusion of procedure codes
- Explore validation in external cohorts

Future applications of NASH cohort

- Merge predicted and observed NASH to permit extension of follow up time (5+ years)
- Larger cohort to allow the examination of rare outcomes

Abbreviations

- AUC– area under (receiver operating characteristic) curve
- BRR– Bayesian risk ratio
- CPT – Current procedural terminology
- ESLD - End stage liver disease
- HCPCS – Healthcare common procedure coding system
- ICD-9-CM /10CM– International Classification of Diseases, 9th revision/10th revision, Clinical Modification
- LASSO – least absolute shrinkage and selection operator
- LOGO CV – leave one group out cross validation method
- NAFLD – nonalcoholic fatty liver disease
- NASH – nonalcoholic steatohepatitis
- PPV – Positive Predictive Value

Presented at the International Liver Congress (EASL), 10-14 April 2019, Vienna, Austria

DOI: 10.3252/pso.eu.ILC2019.2019
Metabolism, alcohol and toxicity
Laura Telep
SAT-331
ILC2019