

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

Laura E. Telep,1 Chi Zhang,2 Stephen Djesjous,1 Robert P. Myers1, Robertino Mera,1 Alan Hubbard,3 Anand P. Chokkalingam1,2

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

**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 versions of the International Classification of Diseases (ICD-9).
- In ICD-10 (adopted in the US on 1 Oct 2015), NASH has a unique diagnosis code

**Aims**

- To use machine learning to identify ICD-9-NASH/NAFLD 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 ICD-10 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 NASH/NAFLD 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 Pharrmatics Plus™ Claims database, including US administrative claims for ~140M patient-lives from 01 Jan 2006 to 30 Mar 2018
  - Patients included in the analysis had a claim for NASH/NAFLD in version 9 of the International Classification of Disease (ICD-9).

- **Data Set for Predictive Algorithm Creation**
  - Every patient included in the data set:
    - had an ICD-9 claim for NASH/NAFLD (571.8)
    - were at least 18 years of age on 1 Oct 2015 (index date)
    - had continuous enrollment during the three years of the ICD-9 era (01 Oct 2013 – 30 Sep 2015)
  - NASH cohort patients (434,177) had a claim for NASH (K75.81) in the first year of the ICD-10 era (01 Oct 2015 – 30 Sep 2016)
  - NASH in ICD-10 (K75.81) was comprised of a random – 50% sample of patients with no claims for ICD-10 NASH in any observation time in the ICD-10 era (starting 01 Oct 2015)

- **Data Set for Predictive Algorithm Creation**
  - Every patient included in the data set:
    - had an ICD-9 claim for NASH/NAFLD (571.8)
    - were at least 18 years of age on 1 Oct 2015 (index date)
    - had continuous enrollment during the three years of the ICD-9 era (01 Oct 2013 – 30 Sep 2015)
  - NASH cohort patients (434,177) had a claim for NASH (K75.81) in the first year of the ICD-10 era (01 Oct 2015 – 30 Sep 2016)
  - NASH in ICD-10 (K75.81) was comprised of a random – 50% sample of patients with no claims for ICD-10 NASH in any observation time in the ICD-10 era (starting 01 Oct 2015)

**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 versions of the International Classification of Diseases (ICD-9).
- In ICD-10 (adopted in the US on 1 Oct 2015), NASH has a unique diagnosis code

**Aims**

- To use machine learning to identify ICD-9-NASH/NAFLD 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 ICD-10 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 NASH/NAFLD 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 Pharrmatics Plus™ Claims database, including US administrative claims for ~140M patient-lives from 01 Jan 2006 to 30 Mar 2018
  - Patients included in the analysis had a claim for NASH/NAFLD in version 9 of the International Classification of Disease (ICD-9).

- **Data Set for Predictive Algorithm Creation**
  - Every patient included in the data set:
    - had an ICD-9 claim for NASH/NAFLD (571.8)
    - were at least 18 years of age on 1 Oct 2015 (index date)
    - had continuous enrollment during the three years of the ICD-9 era (01 Oct 2013 – 30 Sep 2015)
  - NASH cohort patients (434,177) had a claim for NASH (K75.81) in the first year of the ICD-10 era (01 Oct 2015 – 30 Sep 2016)
  - The NASH in ICD-10 (K75.81) was comprised of a random – 50% sample of patients with no claims for ICD-10 NASH in any observation time in the ICD-10 era (starting 01 Oct 2015)

**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 versions of the International Classification of Diseases (ICD-9).
- In ICD-10 (adopted in the US on 1 Oct 2015), NASH has a unique diagnosis code

**Aims**

- To use machine learning to identify ICD-9-NASH/NAFLD 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 ICD-10 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 NASH/NAFLD 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 Pharrmatics Plus™ Claims database, including US administrative claims for ~140M patient-lives from 01 Jan 2006 to 30 Mar 2018
  - Patients included in the analysis had a claim for NASH/NAFLD in version 9 of the International Classification of Disease (ICD-9).

- **Data Set for Predictive Algorithm Creation**
  - Every patient included in the data set:
    - had an ICD-9 claim for NASH/NAFLD (571.8)
    - were at least 18 years of age on 1 Oct 2015 (index date)
    - had continuous enrollment during the three years of the ICD-9 era (01 Oct 2013 – 30 Sep 2015)
  - NASH cohort patients (434,177) had a claim for NASH (K75.81) in the first year of the ICD-10 era (01 Oct 2015 – 30 Sep 2016)
  - The NASH in ICD-10 (K75.81) was comprised of a random – 50% sample of patients with no claims for ICD-10 NASH in any observation time in the ICD-10 era (starting 01 Oct 2015)