Machine Learned Decision Tree for Diagnosis of ASMD Among Patients with Unexplained ILD

Imre Noth[1], Francesco Bonella[2], Wim A. Wuyts[3], Pauline Guilmin[4], Margaux Törnqvist[4], Stefaan Sansen[5], Alexandra Dumitriu[6], Neha Shah[6]*, Maja Gasparic[7], Martin Montmerle[4]

[1]Pulmonary and Critical Care Division, University of Virginia School of Medicine, Charlottesville, VA; [2]Center for Interstitial and Rare Lung Disease Unit, University of Duisburg-Essen, Ruhrlandklinik, Essen, Germany; [3]Unit for Interstitial Lung Diseases, Department of Respiratory Medicine, University Hospitals Leuven, Leuven, Belgium; [4]Quinten Health, Paris, France; [5]Sanofi, Diegem, Belgium; [6]Sanofi, Cambridge, MA, USA; [7]Sanofi, Amsterdam, The Netherlands
**Former employee of Sanofi, Cambridge, MA, USA*

INTRODUCTION

- Acid sphingomyelinase deficiency (ASMD), also known as Niemann-Pick disease types A, A/B and B, is a rare, progressive, and potentially life-threatening lysosomal storage disease caused by biallelic pathogenic variants in *SMPD1*, the gene encoding acid sphingomyelinase (ASM). ASMD leads to accumulation of sphingomyelin in spleen, lungs, liver, and other organs, as well as brain in more severe phenotypes[1].
- The most common clinical manifestations of ASMD include interstitial lung disease (ILD) (>80% of patients) and splenomegaly (>90% of patients)[1].
- Olipudase alfa (recombinant human ASM) is the only approved disease-specific treatment for the non-central nervous system manifestations of ASMD in adults and children[2].
- Diagnosis of rare genetic diseases is challenging because of overlapping symptoms with more common diseases and low disease awareness; a delayed diagnosis and misdiagnosis are common and hinder appropriate management and monitoring of disease manifestations[3-5].
- After initial clinical suspicion, ASMD is diagnosed by enzyme and genetic testing in leukocytes, dried blood spots or cultured fibroblasts[6].
- As ~25% of ILD patients remain unclassified, we hypothesise that some ASMD patients may be identified among patients with unexplained ILD[7].
- Machine learning approaches, such as data-learnt decision trees, can be applied to large datasets, such as electronic health records (EHRs), to flag potential rare disease patients[8].

OBJECTIVE

- To develop a diagnostic decision tree algorithm using clinical and laboratory traits associated with ASMD to facilitate identification of ASMD patients among patients with unexplained ILD, by using machine learning

METHODS

Data source

- EHR data from Optum's de-identified Integrated Claims-Clinical dataset (2007 Q1 – 2020 Q4) were used to select the study population, to extract and derive patient features, and to perform analyses.
- These data integrate multiple EHRs from across the continuum of care, both inpatient and ambulatory.

Cohort creation

- ASMD and control cohorts were created by using two types of information: diagnoses (ICD codes) and provider notes.
- ASMD cohort (n=31):
  - Patients with diagnosis of ASMD or Niemann-Pick disease types A, A/B, B, or unspecified, and pulmonary symptom(s) were included.
  - Patients with any diagnosis of Niemann-Pick disease types C or D were excluded.
  - Patients with incoherent disease journey timelines (e.g., date of death, first or last date of activity before date of ASMD diagnosis) or missing gender information were excluded.
- Control cohort (n=620):
  - Random sampling was used to select 20 unique controls with pulmonary symptoms matched to each ASMD patient in terms of age, gender, region, enrolment duration, and first active year group.

Selection of clinical variables

- Clinical variables were initially selected to capture hallmark characteristics of ASMD available in EHR data; this initial set was enriched with variables prioritized using statistical methods.
- For the model development, patient characteristics were extracted from three data sources: diagnoses, procedures, and laboratory measurements.
- Three types of features were derived from the extracted patient characteristics: symptoms (binary), symptom groups (binary) and laboratory measurements (continuous).

Model training and evaluation

- A decision tree was trained on the considered clinical variables to differentiate ASMD patients from control patients with pulmonary manifestations.
- Optimal hyperparameters for the algorithm were selected using a cross-validation approach on 5 folds.
- The algorithm was internally validated on the ASMD and control cohorts.

Application of model to young unexplained ILD population

- The trained decision tree algorithm was applied to a patient population with unexplained ILD available from the same Optum data source.
- The ILD unexplained cohort (N=270,549) was defined using ICD 10 codes J84.9, J84.10 or ICD9 codes 5169 and then filtered on younger patients (age≤50 years) to obtain a young unexplained ILD cohort (N=35,930).

Descriptive statistics

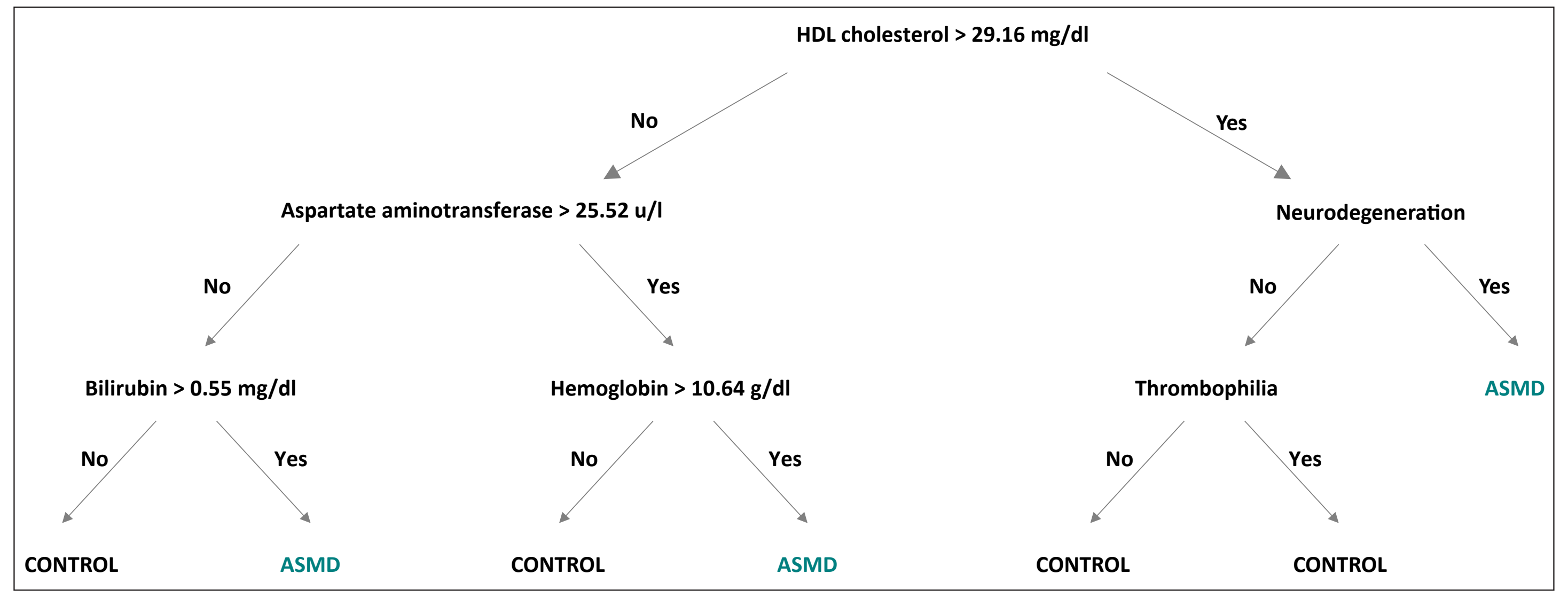
- Descriptive analyses were used for demographics, clinical characteristics, and laboratory measurements to describe and compare relevant cohorts. The following metrics were calculated:
  - Mean and standard mean deviation (SD) of the number of clinical characteristics by patient
  - Prevalence (%): frequency of occurrence of a symptom in each cohort

RESULTS

Machine learned decision tree

- Given the available EHR data from Optum's de-identified Integrated Claims-Clinical dataset (2007 Q1 – 2020 Q4), the ASMD cohort with pulmonary manifestations used for model training was enriched with 199 clinical characteristics and 11 laboratory measurements.
- The most prevalent manifestations among the ASMD patients with pulmonary involvement included splenomegaly (55%), lung infection (48%), dyspnea (48%), hyperlipidemia (45%), chest pain (42%), anemia (39%), GERD (35%), abdominal pain (35%), hepatomegaly (32%) and thrombocytopenia (32%).
- The machine learned decision tree for ASMD highlighted four laboratory measurements (HDL cholesterol, aspartate aminotransferase, bilirubin, and hemoglobin) and one symptom (neurodegeneration).
- The generated decision tree is presented in **Figure 1** and the associated rules are displayed in **Table 1**.

Figure 1. Machine learned decision tree*



HDL cholesterol > 29.16 mg/dl
No
Aspartate aminotransferase > 25.52 u/l
No
Bilirubin > 0.55 mg/dl
No
CONTROL
Yes
ASMD
Yes
Hemoglobin > 10.64 g/dl
No
CONTROL
Yes
ASMD
Yes
Neurodegeneration
No
Thrombophilia
No
CONTROL
Yes
CONTROL
Yes
ASMD

*Model hyperparameters: maximal depth=3, minimum samples for splitting leaves=5, class weights={0:1, 1:1}

- The algorithm distinguished ASMD vs. matched controls, with a sensitivity of ~80% and a specificity >99% (**Table 1**).

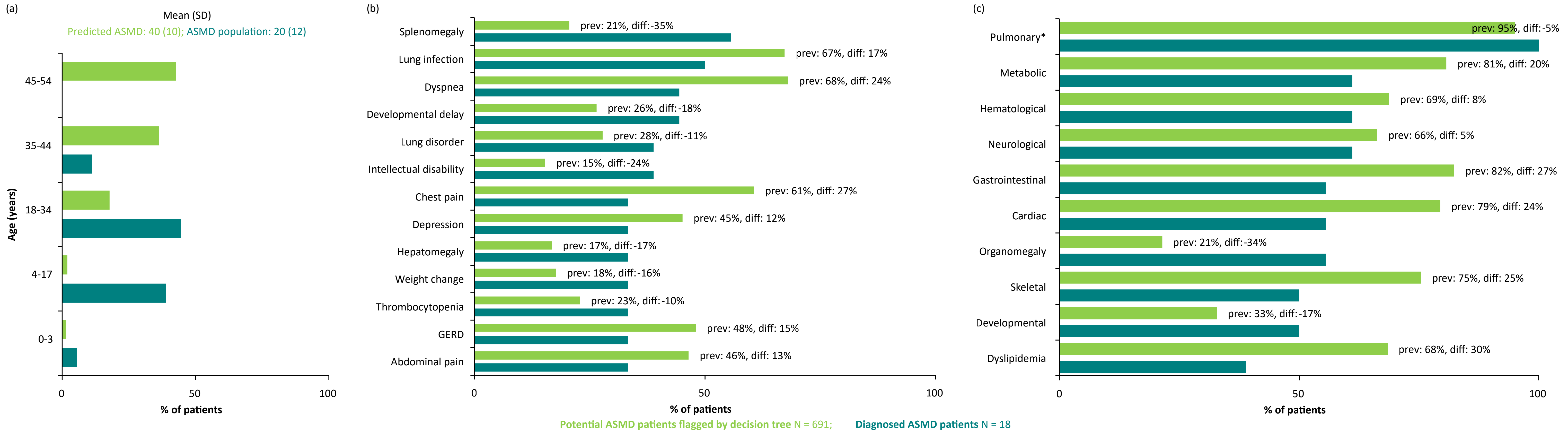
Table 1: Decision tree rule description for train-test set

Rule description	Train-test set (N = 651)	
	ASMD patients (N = 31)	Control patients (N = 620)
	N (%)	N (%)
HDL cholesterol ≤ 29.16 mg/dl **and** aspartate aminotransferase > 25.52 u/l **and** hemoglobin > 10.64 g/dl	20 (64.5%)	0 (0.0%)
HDL cholesterol ≤ 29.16 mg/dl **and** aspartate aminotransferase ≤ 25.52 u/l **and** bilirubin > 0.55 mg/dl	3 (9.7%)	1 (0.2%)
HDL cholesterol > 29.16 mg/dl **and** neurodegeneration = Yes	2 (6.5%)	1 (0.2%)
Total tree (union of rules)	25 (80.7%)	2 (0.4%)

Application of the algorithm on the young unexplained ILD population

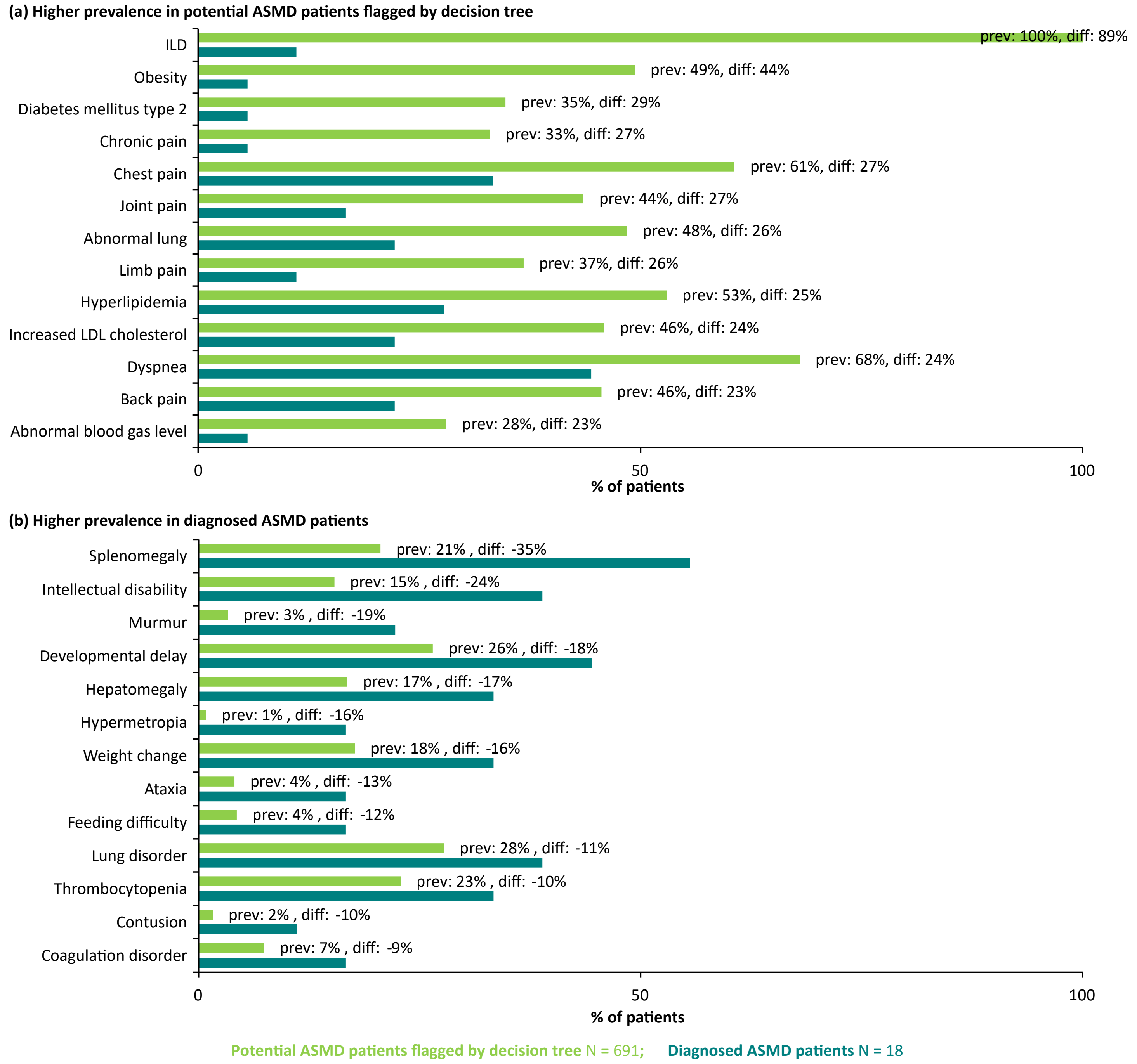
- Application to a young unexplained ILD cohort ≤50 years (N=35,930) flagged 691 (1.9%) potential ASMD patients, which represents a reasonably small proportion of the population to proceed with enzyme and genetic testing to confirm ASMD diagnosis.
- Potential ASMD patients flagged by the decision tree were on an average older than diagnosed ASMD patients < 50 years (n=18) and had less prevalent splenomegaly and organomegaly overall. Lung infections, dyspnea and chest pain were more common, however pulmonary manifestations as a symptom group were similarly reported in both cohorts (**Figure 2**).
- Top clinical differences between the two cohorts (diagnosed ASMD vs. decision tree flagged ASMD) are displayed in **Figure 3**. They show over-representation of metabolic syndrome signs in decision tree flagged ASMD patients, such as diabetes mellitus type 2, obesity, hyperlipidemia, increased LDL cholesterol, which can be attributed to the older age in this cohort.

Figure 2. Age and ASMD clinical characteristics in diagnosed ASMD patients vs potential ASMD patients flagged by decision tree
(a) Age (years) (b) Top ASMD features (c) Top ASMD feature groups



*Without ILD
diff, difference; GERD, Gastroesophageal Reflux Disease; prev, prevalence.

Figure 3. Top differences in clinical characteristics between diagnosed ASMD patients and potential ASMD patients flagged by decision tree



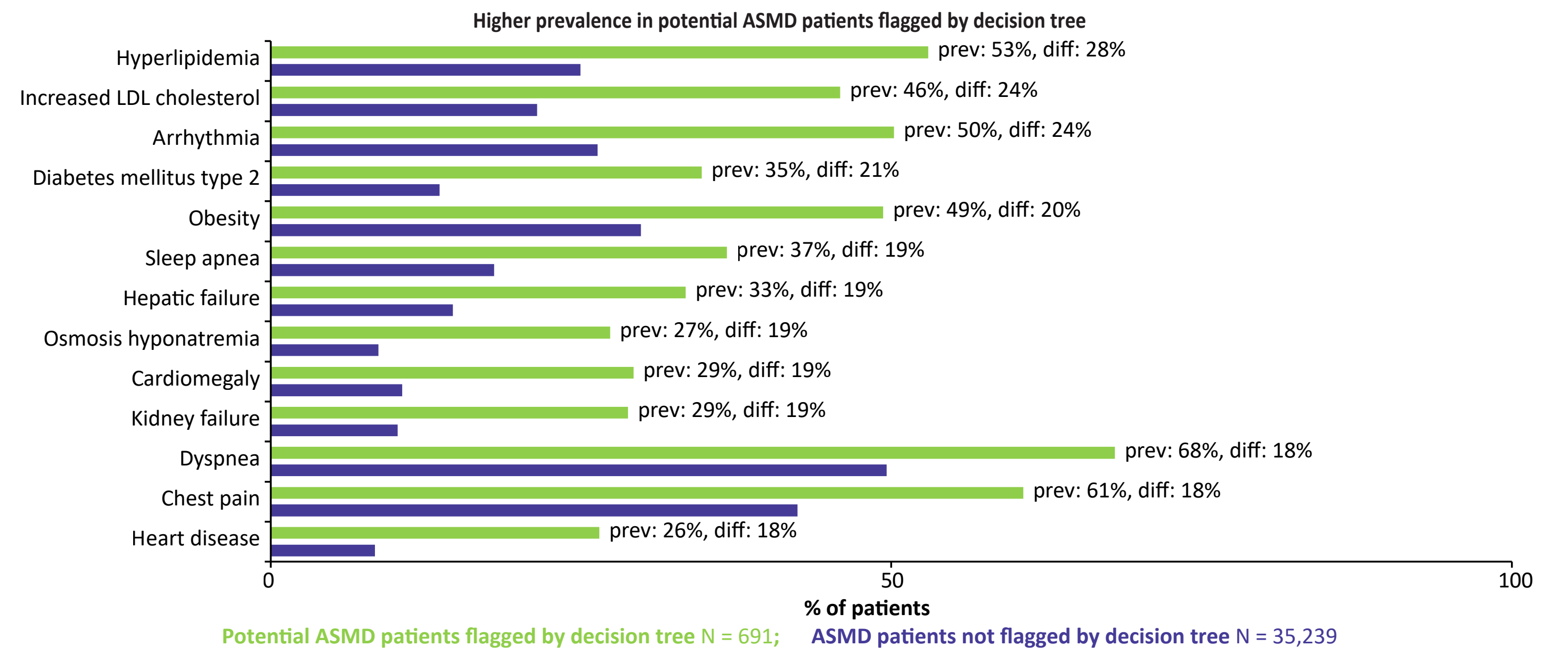
(a) Higher prevalence in potential ASMD patients flagged by decision tree

(b) Higher prevalence in diagnosed ASMD patients

diff, difference; ILD, Interstitial Lung Disease; prev, prevalence.

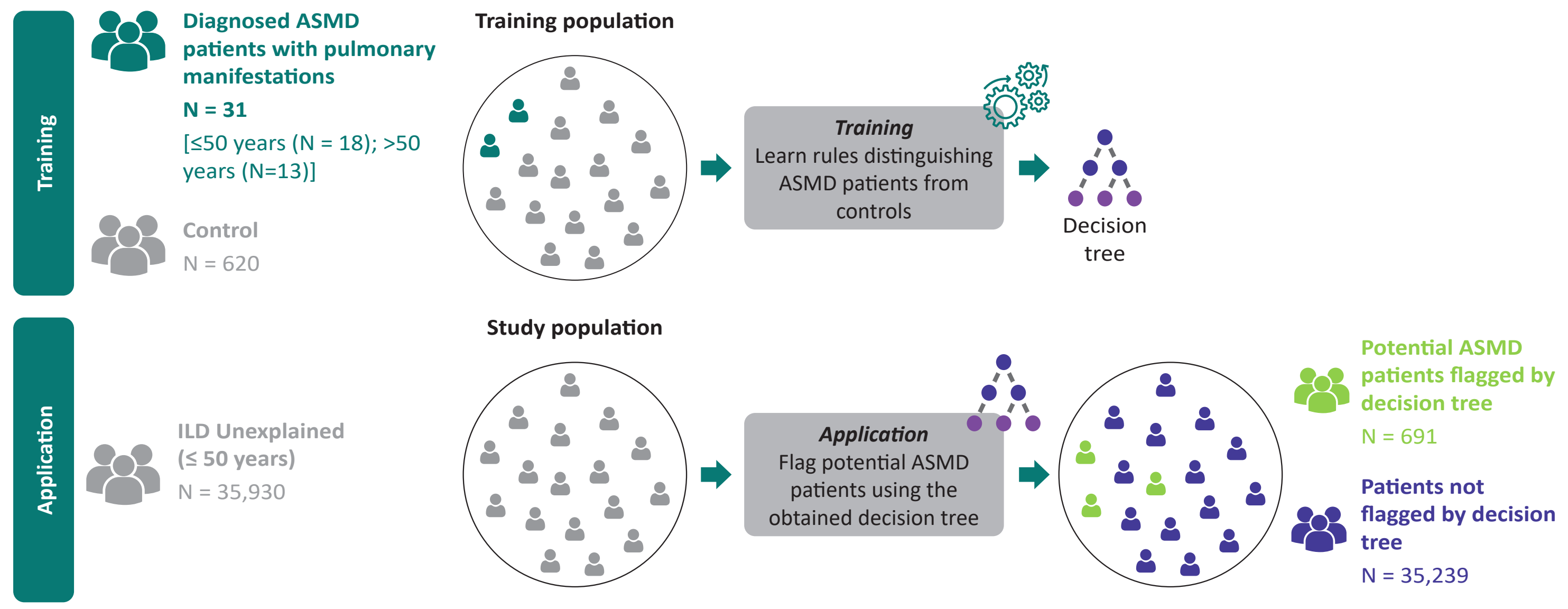
- Potential ASMD patients that were flagged by the decision tree have more signs associated with the metabolic syndrome, cardiac manifestations and kidney failure, some of which are ASMD manifestations described in the literature[1,3], compared to patients with unexplained ILD <50 years, who were not flagged by the decision tree (**Figure 4**).

Figure 4. Top differences between potential ASMD patients flagged by decision tree and those that were not flagged



- An overview of the project flow and associated populations is shown in **Figure 5**.

Figure 5. Training and application of machine learned decision tree



Training
Diagnosed ASMD patients with pulmonary manifestations
N = 31
[≤50 years (N = 18); >50 years (N=13)]
Control
N = 620
Training population
Training
Learn rules distinguishing ASMD patients from controls
Decision tree

Application
ILD Unexplained (≤ 50 years)
N = 35,930
Study population
Application
Flag potential ASMD patients using the obtained decision tree
Potential ASMD patients flagged by decision tree
N = 691
Patients not flagged by decision tree
N = 35,239

CONCLUSIONS

- The machine learned algorithm was able to capture ASMD types of patients from EHR data with great sensitivity and to flag a reasonably small number (<2%) of potential ASMD patients in the unexplained ILD cohort < 50 years.
- This algorithm may enhance early diagnosis of ASMD, though validation is required.
- Differences with regards to age and clinical characteristics observed between cohorts may help to further filter out potential patients with ASMD, who may then undergo enzymatic and genetic testing to confirm ASMD diagnosis.

ACKNOWLEDGMENTS

- The study is sponsored by Sanofi and the medical writing support was provided by Deepa Warrier from Sanofi.
- The text, figures, and tables in this presentation cannot be reproduced without the explicit permission of the authors and Sanofi.

DISCLOSURES

- **Imre Noth** has received grants from the National Heart, Lung and Blood Institute; and consulting fees from Boehringer Ingelheim, Genentech and Confo.
- **Francesco Bonella** reports lecture honoraria from Boehringer Ingelheim, Fujirebio, Galapagos NV, Roche; travel support from Boehringer Ingelheim, Roche; participation on advisory boards with Boehringer Ingelheim, Bristol Myers Squibb, Fujirebio, Galapagos NV, GlaxoSmithKline, Roche, Takeda.
- **Wim A. Wuyts** reports grants received to institution or contracts from Roche, Boehringer Ingelheim, Galapagos, the NIH and Fund Scientific Research Flanders.
- **Pauline Guilmin, Margaux Törnqvist** and **Martin Montmerle** are employees of Quinten Health and serve as consultants to Sanofi.
- **Stefaan Sansen, Alexandra Dumitriu, Neha Shah** and **Maja Gasparic** are/were employees of Sanofi and may hold stock/stock options.

REFERENCES

1. Wasserstein M. et al. *Mol Genet Metab*. 2019 Feb;126(2):98-105.
2. Lachmann RH. Et al. *Orphanet J Rare Dis*. 2023 Apr 25;18(1):94.
3. McGovern MM. et al. *Genet Med*. 2017 Sep;19(9):967-974.
4. Wang RY. et al. *Genet Med*. 2011 May;13(5):457-84.
5. Toussi T. et al. *Can Med Assoc J*. 1974 Sep 21;111(6):556, 559-60.
6. Pinto C et al. *Int J Mol Sci*. 2021 Nov 28;22(23):12870.
7. Traila D. et al. *J Int Med Res*. 2018 Jan;46(1):448-456.
8. Davenport T. et al. *Future Healthc J*. 2019 Jun;6(2):94-98.