Utilizing Tokenization to Integrate Three Data Sources in Rare Disease Research: Muscular Dystrophy - Longitudinal Integrated Claims (MD-LINC)

Bryan Innis,[1] Sourav Santra,[1] Shane Hornibrook,[1] Jacko Logan,[1] Richard Baxter,[1] Katherine Gooch[1]

[1]Sarepta Therapeutics, Inc., Cambridge, MA, USA

Background

- Muscular dystrophy (MD) refers to a group of genetic diseases that cause progressive weakness and degeneration of skeletal muscles[1]
- These disorders (of which there are more than 30) vary in the age of onset, severity, and pattern of the affected muscles[1]
- Duchenne muscular dystrophy (DMD) is an X-linked, degenerative neuromuscular disease caused by genetic mutations in the *DMD* gene that affects approximately 7.1 per 100,000 males globally[2] and 9,000 to 12,000 males in the US[3]
- The *International Classification of Diseases (ICD)* code for DMD (G71.01) includes both DMD and Becker muscular dystrophy (BMD); therefore, identification of patients with DMD from administrative claims data is challenging[4,5]
- Despite significant research, there is limited understanding of the demographic and clinical characteristics, health care utilization, diagnostic and treatment journey, and overall survival among patients with MD
- Validating algorithms with genetic laboratory results may help provide a more robust method to accurately identify and differentiate between patients with DMD and BMD in claims data[1]
- The development of a tokenized data set that includes genetic laboratory results alongside administrative claims and mortality data may help clinicians distinguish between MD types and fill existing knowledge gaps, offering insight into the care and long-term outcomes of patients with MD

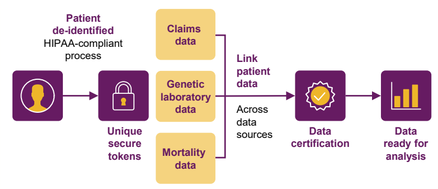
Objective

To develop a tokenized data set, Muscular Dystrophy - Longitudinal Integrated Claims (MD-LINC), of patients with MD by linking administrative, mortality, and genetic laboratory data

Methods

- This research study adopted a retrospective cohort design to develop a tokenized health care data set of patients with MD
- Datavant© proprietary tokenization software was used to integrate 3 distinct data sources: Inovalon closed claims, Datavant Mortality, and consolidated genetic laboratory data
- A remediation protocol provided guidance for censoring data elements not pertinent for research purposes, enabling reliable identification and long-term tracking of patients with MD within a real-world context
- Privacy Hub completed an expert attestation to certify the de-identified data, per the Health Insurance Portability and Accountability Act (HIPAA) Expert Determination method [Section 164.514(b) (1) of the HIPAA Privacy Rule] (**Figure 1**)
- All patients with at least one administrative claim with a diagnosis of MD or of a non-specific muscle disorder (*ICD-9*: 359.1, 359.29, 359.89, 359.9; *ICD-10*: G71.0x, G71.8, G71.9, G71.19) in the claims data were identified and flagged as patients with MD (**Figure 1**)
- When laboratory results were available to determine *DMD* genetic status, differentiation between DMD-like and BMD-like genotypes was made based on mutation type and the published literature
- The index date was the first diagnosis of MD between July 1, 2012, and September 30, 2024
- Descriptive statistics were generated for the source and resulting data sets
  - All results were reported as an approximate value

Figure 1 *Data Tokenization*



HIPAA, Health Insurance Portability and Accountability Act; MD, muscular dystrophy.

Results

- Of more than 200,000 individuals, approximately 7% (~15,000) had a genetic test for MD during the study period, creating the MD-LINC data set
- Most of the patients included were male (57.5%) and almost one-third were White (31.7%)
- Over 40% had commercial insurance with a median duration of 3.3 years of medical and pharmacy insurance (**Table 1**)
- In patients with a laboratory result, positive *DMD* genetic status was recorded in 12.9% of patients, with 8.2% and 4.8% having DMD-like and BMD-like genotype–phenotypes, respectively (**Table 2**)
- The developed MD-LINC data set will provide a basis for future research and different research questions to be asked (**Table 3**)
  - For example, cohort 4 represents ~15,000 patients with genetic information and health insurance claims data (treated and untreated) available
  - Cohort 5 is the "fully" linked data set, containing records across the data set for all patients who appear in the Datavant Mortality database, including month of death

Table 1 *Study Population*

Parameter, %	Patient in genetic laboratory database and Inovalon therapeutic cohort (N=~15,000)
Sex	
Male	57.5
Female	42.5
Race	
White	31.7
Black or African American	6.6
Asian or Pacific Islander	2.2
Other	5.2
Unknown	46.7
Ethnicity	
Hispanic or Latino	7.6
Payer group	
Commercial	43.1
Medicaid	34.8
Medicare Advantage	4.4
Unknown	1.8

Table 2 *Genetic Status*

Parameter, %	Patient in genetic laboratory database and in Inovalon therapeutic cohort (N=~15,000)
DMD genetic status	
Negative	86.1
Positive	12.9
Uncertain	0.9
Inferred genotype–phenotype	
DMD-like	8.2
BMD-like	4.8

BMD, Becker muscular dystrophy; DMD, Duchenne muscular dystrophy.

- To demonstrate application, the MD-LINC data set was used to develop a patient identification algorithm (described in the Application in Action section) in a DMD study[6]

Table 3 *MD-LINC Data Set*

Cohort		n
1	Present in genetic laboratory database OR Inovalon therapeutic cohort	228,000
2	Present in genetic laboratory database	33,000
3	Present in Inovalon therapeutic cohort	210,000
4	Present in genetic laboratory database AND in Inovalon therapeutic cohort	15,000
5	Present in genetic laboratory database AND in Inovalon therapeutic cohort AND present in Datavant Mortality database	<300

Application in Action

- Previously, the linked data set provided the foundation for patient identification algorithm development[6]
- Three algorithms including a random forest machine learning approach (broad, narrow, and restrictive; **Figure 2**) were tested
- The algorithms were tested using data from insured US patients from the Inovalon closed claims database
- A reference standard for algorithm performance testing was determined using genetic data testing

Figure 2 *Algorithms' Criteria for Classifying Patients as Having DMD*[4]

Broad DMD definition
- Male
- Age ≤40 years at first DMD/BMD diagnosis code
- ≥2 claims with DMD/BMD diagnosis code

Narrow DMD definition
- Met **broad DMD definition**
- Had ≥1 of the following:
  - Prescription for glucocorticoids[a] at any time
  - Claim for exon-skipping therapy[b] or gene therapy[c]
  - Evidence of LOA by age 12
  - Evidence of ventilation support or dependence at any age
- Patients aged ≥30 years were required to have evidence of ventilation support or dependence on or before their 30th year

Restrictive DMD definition
- Met **narrow DMD definition**
- Patients aged ≥20 years were required to have evidence of ventilation support or dependence on or before their 20th year

[a]Glucocorticoids included betamethasone, budesonide, cortisone, deflazacort, dexamethasone, hydrocortisone, methylprednisolone, prednisolone, prednisone, or triamcinolone. [b]Exon-skipping therapy included casimersen, eteplirsen, golodirsen, or viltolarsen. [c]Gene therapy included delandistrogene moxeparvovec-rokl.
BMD, Becker muscular dystrophy; DMD, Duchenne muscular dystrophy; LOA, loss of ambulation.

- Patients were considered to have a DMD-like genotype if they had a mutation in the *DMD* gene resulting in a frameshift and/or a stop codon
- Patients were considered negative for DMD if they did not have a mutation in the *DMD* gene or had a mutation in the *DMD* gene not resulting in a frameshift and/or stop codon (referred to as BMD-like)
- Patients with an uncertain test result for *DMD* mutations were excluded
- Algorithm performance against the reference standard was evaluated using confusion matrices to calculate sensitivity, specificity, positive predicted value (PPV), and negative predictive value (NPV)
- None of the 3 published algorithms reached the predetermined PPV >80% or Cohen's kappa statistic >0.60[6]
- Results of a random forest classification analysis indicated that the removal of loss of ambulation (LOA) by age 12 from the narrow and restrictive algorithms may improve the model's accuracy
- After removal of LOA by age 12, the two algorithms (narrow, and restrictive) were able to sufficiently identify and distinguish between patients with DMD and BMD (**Figure 3**)

Figure 3 *Modified Previously Published Claims-Based Algorithms*

A. Confusion matrices of modified algorithms

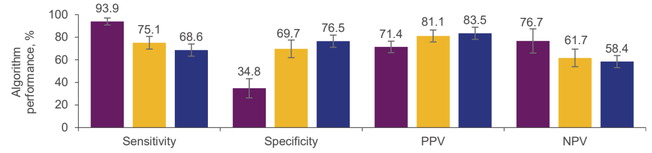
Modified narrow DMD algorithm

Algorithm predictions	Reference standard: Has DMD[a]	Reference standard: No DMD[b]
Has DMD[a]	172 (47.6%)	40 (11.1%)
No DMD[b]	57 (15.8%)	92 (25.5%)

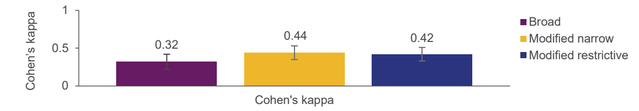
Modified restrictive DMD algorithm

Algorithm predictions	Reference standard: Has DMD[a]	Reference standard: No DMD[b]
Has DMD[a]	157 (43.5%)	31 (8.6%)
No DMD[b]	72 (19.9%)	101 (28.0%)

B. Summary of algorithm performance



C. Summary of Cohen's kappa results



[a]"Has DMD" per algorithm prediction indicates patients were classified as having DMD based on algorithm criteria (**Figure 2**). [b]"No DMD" per reference standard indicates phenotype is BMD-like, or DMD genetic test result was uncertain/negative; "No DMD" per algorithm prediction indicates patients were classified as not having DMD based on algorithm criteria (**Figure 2**).
Bars represent 95% CI.
BMD, Becker muscular dystrophy; DMD, Duchenne muscular dystrophy; NPV, negative predictive value; PPV, positive predictive value.

- The Cohen's kappa statistic was <0.60 in the broad, modified narrow, and modified restrictive algorithms, with modest improvement observed for the modified narrow algorithm (**Figure 3C**)
- Overall, the modified narrow algorithm had the best balance of PPV and Cohen's kappa and should be used in future Inovalon claims studies to identify cohorts of patients with DMD
  - However, the low NPV and Cohen's kappa statistic results indicate that some patients who had DMD were missed; the algorithm should be refined and maintained as new treatments become available

Conclusions

- The ability to link 3 data sources to create the MD-LINC data set provides a unique opportunity to conduct health economics and outcomes research as well as real-world evidence research regarding rare diseases that has not been possible previously
  - In the future, applications of interest could incorporate patient-, caregiver-, or clinician-reported outcomes, phase 4 studies, and patient registries
- In an example study, application of the MD-LINC data set enabled the utilization of machine learning to develop an algorithm to identify patients with true DMD (distinguished from BMD) that was not previously possible, supporting improved identification of future study populations and matched controls
  - This will help improve the accuracy of patient identification and fill existing knowledge gaps, offering insight into the care and long-term outcomes of patients with MD

Acknowledgments and Disclosures

Acknowledgments: This study was funded by Sarepta Therapeutics, Inc., Cambridge, MA, USA. Editorial support was provided by Barrie Anthony, PhD, of Envision 90TEN, an Envision Medical Communications agency, a part of Envision Pharma Group, in accordance with Good Publication Practice (GPP) 2022 guidelines (https://www.ismpp.org/gpp-2022) and was funded by Sarepta Therapeutics, Inc., Cambridge, MA, USA. **Disclosures: BI, SS, SH, JL, RB,** and **KG**: All authors are employees of Sarepta Therapeutics, Inc., and may own stock/options in the company.

References

1. Wilson K, et al. *Toxicol Pathol*. 2017;45(7):961-976. **2.** Crisafulli S, et al. *Orphanet J Rare Dis*. 2020;15(1):141. **3.** Mendell JR, et al. *J Neuromuscul Dis*. 2021;8(4):469-479. **4.** Schrader R, et al. *J Manag Care Spec Pharm*. 2023;29(9):1033-1044. **5.** Gooch KL, et al. *Adv Ther*. 2024;41(9):3615-3632. **6.** Grabich S, et al. Poster presented at: ISPE 2025; August 22-26, 2025; Washington, DC. Poster C-201.

PRESENTED AT ISPOR 2026
MAY 17-20, 2026
PHILADELPHIA, PA