Available online at www.sciencedirect.com

ScienceDirect

Neuromuscular Disorders 32 (2022) 271–283

www.elsevier.com/locate/nmd

# Real-world and natural history data for drug evaluation in Duchenne muscular dystrophy: suitability of the North Star Ambulatory Assessment for comparisons with external controls

Francesco Muntoni[a], James Signorovitch[b,c,*], Gautam Sajeev[b], Nathalie Goemans[d], Brenda Wong[e], Cuixia Tian[f], Eugenio Mercuri[g], Nicolae Done[b], Hallee Wong[b], Jackson Moss[b], Zhiwen Yao[b], Susan J. Ward[c], Adnan Manzur[a], Laurent Servais[h], Erik H. Niks[i], Volker Straub[j], Imelda JM de Groot[k], Craig McDonald[l], The North Star Clinical Network, PRO-DMD-01 Study, The Association Française contre les Myopathies (AFM), The DMD Italian Group, and The Collaborative Trajectory Analysis Project (cTAP)

[a] *Dubowitz Neuromuscular Centre, NIHR Great Ormond Street Hospital Biomedical Research Centre, Great Ormond Street Institute of Child Health, University College London, & Great Ormond Street Hospital Trust, London, UK*

[b] *Analysis Group, Inc., Boston, Massachusetts, USA*

[c] *The Collaborative Trajectory Analysis Project, Cambridge, Massachusetts, USA*

[d] *University Hospitals Leuven, Child Neurology, Leuven, Belgium*

[e] *Department of Pediatrics, University of Massachusetts Medical School, Worcester, Massachusetts, USA*

[f] *Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio & College of Medicine, University of Cincinnati, Cincinnati, Ohio, USA*

[g] *Department of Pediatric Neurology, Fondazione Policlinico Gemelli IRCCS, Catholic University, Rome, Italy*

[h] *UK Oxford Neuromuscular Center, Department of Paediatrics, University of Oxford, UK and Neuromuscular Center of Liège, Division of Paediatrics, CHU and University of Liège, Belgium*

[i] *Leiden University Medical Centre, Leiden, the Netherlands*

[j] *John Walton Muscular Dystrophy Research Centre, Newcastle University, Newcastle, UK*

[k] *Radboud University Nijmegen Medical Center, Donders Centre of Neuroscience, Department of Rehabilitation, Nijmegen, the Netherlands*

[l] *Department of Physical Medicine and Rehabilitation, and Pediatrics, University of California, Davis, Sacramento, California, USA*

Received 16 August 2021; received in revised form 15 February 2022; accepted 18 February 2022

**Abstract**

Using external controls based on real-world or natural history data (RWD/NHD) for drug evaluations in Duchenne muscular dystrophy (DMD) is appealing given the challenges of enrolling placebo-controlled trials, especially for multi-year trials. Comparisons to external controls, however, face risks of bias due to differences in outcomes between trial and RWD/NHD settings. To assess this bias empirically, we conducted a multi-institution study comparing mean 48-week changes in North Star Ambulatory Assessment (NSAA) total score between trial placebo arms and RWD/NHD sources, with and without adjustment for baseline prognostic factors. Analyses used data from three placebo arms (235 48-week intervals, $N = 235$ patients) and three RWD/NHD sources (348 intervals, $N = 202$ patients). Differences in mean ΔNSAA between placebo arms and RWD/NHD sources were small before adjustment (-1.2 units, 95% CI: [-2.0 -0.5]) and were attenuated and no longer statistically significant after adjustment (0.1 units (95% CI: [-0.6, 0.8]). Results were similar whether adjusting using multivariable regression or propensity score matching. This consistency in ΔNSAA between trial placebo arms and RWD/NHD sources accords with prior findings for the six-minute walk distance, provides a well-validated framework for baseline adjustment of prognostic factors, and supports the suitability of RWD/NHD external controls for drug evaluations in ambulatory DMD.



*Keywords:* Clinical trials; Drug evaluation; Duchenne muscular dystrophy; External controls; Natural history data; Real-world data.

* Corresponding author.
*E-mail address:* James.signorovitch@analysisgroup.com (J. Signorovitch).

https://doi.org/10.1016/j.nmd.2022.02.009

## 1. Introduction

Uses of external controls drawn from real-world data and natural history data (RWD/NHD) are of high interest for drug evaluations in Duchenne muscular dystrophy (DMD), and have informed drug evaluation in other rare conditions [1–4]. While randomized, placebo-controlled trials are the gold standard for measuring drug efficacy and safety, a number of factors complicate the methodological, practical, and ethical considerations for placebo controls in DMD trials, especially for longer periods of follow-up beyond 48 weeks. Variability in rates of disease progression across individuals reduces power to detect meaningful treatment effects with readily achievable sample sizes. Safe and effective blinding to treatment assignment is not always feasible with invasive administration procedures or tell-tale reactions to active therapy [5,6]. Small sample sizes admit baseline imbalances despite randomization [7]. Enrollment challenges can arise from the rarity of the disease, the need for longer-term follow-up, and the wealth of promising investigational agents under clinical development [8]. These challenges can become even more pronounced for trials targeting genetic subtypes of DMD [8]. The ethics of assigning patients to long-term, multi-year placebo arms can also be questioned, due to the risks of both extended exposure to potentially ineffective therapy and denial of opportunities to participate in alternative studies [9]. Finally, even when randomized placebo arms are feasible, longer-term follow-up from open-label extension studies, or post-market studies, without placebo controls, may be needed to establish long-term clinical benefit [10]. In all of these situations, use of external controls from RWD/NHD sources holds promise for contextualizing trial outcomes, providing supportive comparative evidence [11,12], augmenting smaller randomized placebo arms, or even serving as a primary comparator group.

The promise of well-designed externally-controlled studies for drug evaluation is further supported by the passage of the 21st Century Cures Act (2016) in the United States, which has generated an increased need for understanding and evaluating appropriate uses of real-world data in regulatory decision making [13,14] and development of frameworks for doing so [15,16]. Representatives of the European Medicines Agency and the organisation for Economic Co-operation and Development have also recognized the importance of RWD/NHD for drug evaluation [17–19]. Against this background, the possible use of RWD/NHD in DMD clinical trials attracts attention as a way to speed trial enrollment and completion, and to enable more patients to access active therapies as opposed to placebos.

At the same time, the prospect of incorporating non-randomized external controls into drug evaluation raises significant and well-founded concerns [20–22]. Without randomization there are numerous avenues through which bias can arise from differences between treatment groups [23,24]. A non-exhaustive list of possibilities is included in Table 1. Particular concerns in DMD include biases arising from differences in baseline ambulatory function and other patient characteristics, dystrophin genotypes and background genetics, levels of physical therapy and other supportive care, and dosing of corticosteroids. In addition, when outcome measures require the assessment of patient performance on functional tests, differences in assessment procedures, training of evaluators, the subject's effort level, and hope and expectation of improvement could all be hypothesized to differentially influence recorded outcomes within versus outside of clinical trial settings [5,8,25,26]. The magnitude of these potential biases, and the extent to which they can be measured or mitigated through pre-specified and well-justified study designs, will determine the suitability of using external controls in DMD.

We therefore investigated the magnitude of these biases in comparisons of 48-week changes in the North Star Ambulatory Assessment (NSAA) between clinical trial settings and RWD/NHD. The NSAA was designed and validated to measure 17 functional components of particular relevance for ambulant DMD that mark disease progression and are important to patients and caregivers [27,28]. The NSAA score has been used as a primary or secondary endpoint in clinical trials enrolling ambulatory boys with DMD [29–32], included in open-label extension studies [33–35], and studied in several natural history studies and real-world databases [27,36]. In assessing the consistency of NSAA outcomes across data sources, the present study builds on our prior study of the consistency of the six-minute walk distance (6MWD) test across data sources in DMD [37]. We also identify prognostic factors for 48-week change in NSAA total score and evaluate different approaches to mitigating the risk of bias by adjusting for these prognostic factors when comparing to external controls.

Table 1
Factors that may differ between clinical trials and external controls and bias comparisons of outcomes*

| |
|---|
| **Population** |
| - Inclusion and exclusion criteria, and baseline characteristics, especially those known or suspected to be associated with outcomes: e.g., age, demographics, vitals, genotype, steroid treatment history, functional performance |
| - Representativeness of the recruitment process (e.g. recruitment sources, burden of participation, consent process, enrollment competition from other studies) |
| - Completeness of follow-up; extent of and reasons for drop-out |
| **Outcomes** |
| - Outcome definitions, adjudication procedures |
| - Assessment process: training and standardization of procedures |
| - Ascertainment process; frequency and timing of assessments |
| **Setting** |
| - Geographies, time periods and associated demographics and background standards of care |
| - Background treatments and care settings (e.g., center of excellence, community care) |
| - Range of non-study treatment options available; concomitant medications or physical therapy |

*A non-exhaustive list of examples.

## 2. Methods

### 2.1. Data sources

Patient-level data from three clinical trial placebo arms and six RWD or NHD sources accessed by the collaborative Trajectory Analysis Project (cTAP) were used in this study (**Tables A.1** and **A.2**). Clinical trial placebo arm data came from three phase 3 trials in ambulatory DMD: the tadalafil DMD trial (patients primarily enrolled and followed from 2013 to 2015) [38], the ACT-DMD trial of ataluren (2013–2015) [35], and the DEMAND III trial of drisapersen (2011–2013) [39], provided to cTAP by Lilly, PTC Therapeutics and CureDuchenne, respectively. Curated RWD came from Universitaire Ziekenhuizen Leuven (2011–2016), the North Star UK database (2005–2015; http://www.northstardmd.com), and Cincinnati Children's Hospital Medical Center (CCHMC) (2004–2016). Collaborators contributing NHD were the DMD Italian Group (2008–2013), the iMDEX study (2012–2018) [40], and the PRO-DMD-01 prospective natural history study (2012–2016) [41]. Data from the iMDEX study were provided to cTAP by University College London on behalf of the Association Française contre les Myopathies (AFM). Data from the PRO-DMD-01 study were provided by CureDuchenne.

### 2.2. Ethics approvals

RWD/NHD sources were approved by ethics committees from each institution (the University Hospitals Leuven [Leuven], Catholic University, Rome [DMD Italian Group], each participating center of the PRO-DMD-01 study, and the institution review board at the CCHMC [IRB #2010–1881]). For the iMDEX study, ethics review boards at the participating institutions approved the study protocol, consent and assent documents.

For use of the North Star UK data, this project followed Caldicott Guardian regulations and information was entered in the database after written informed consent was obtained from patients' parents. Only anonymous, de-identified data were analyzed. All clinical investigations were conducted according to the principles expressed in the Declaration of Helsinki, following Caldicott Guardian approval.

For all data sources, informed consent/assent was obtained for each participant or caregiver as appropriate before the study procedures were conducted.

### 2.3. Study measures

#### 2.3.1. Outcome assessments

The primary outcome in this study was 48-week change in the NSAA total score (ΔNSAA). The NSAA consists of 17 activities and was developed to evaluate changes in motor ability in ambulatory DMD. Patients' performance on each activity is scored by trained clinical staff as either 0 (unable to perform independently), 1 (performs activity using a modified method but is able to complete independently) or 2 (able to perform independently without modification). The NSAA total score is the sum of scores across all activities and ranges from 0 to 34, with higher scores indicating better function. Details of administration of the NSAA in each center are summarized in the Appendix (**Table A.3**).

In the tadalafil, ACT DMD and DEMAND III trials, patients had outcome assessments at regular intervals over the duration of the 48-week trial. Final assessments occurred at 48 weeks and change in NSAA total score between the baseline and week 48 visits were calculated. RWD/NHD sources, in contrast, had different visit frequencies reflecting real-world care or different protocol specifications (as in the case of PRO-DMD-01, which included follow-up visits every 6 months). To facilitate comparison to ΔNSAA assessed in trials, changes in NSAA in RWD/NHD were calculated based on pairs of visits separated by approximately 48 weeks (between 9 and 13 months). ΔNSAA was then linearly rescaled to estimate 48-week changes. If a patient lost ambulation prior to his endpoint visit, or if a linear re-scaling of his change in NSAA resulted in a projected NSAA at 48 weeks of less than zero, the patient was assumed to have lost the ability to complete the NSAA by week 48 and their ΔNSAA was set to the negative of their baseline NSAA.

#### 2.3.2. Potential prognostic factors

The availability of prognostic factors by data source is summarized in **Table A.4**. Age, height, weight, type of steroid (deflazacort or prednisone), and baseline NSAA total score were available in all data sources except the DMD Italian Group and are known to be prognostic of changes in ambulatory function [20]. The availability of additional known prognostic factors varied across data sources. In the tadalafil, ACT DMD and DEMAND III trial placebo arms, and in Leuven, PRO-DMD-01 and iMDEX, timed rise from floor, 4-stair climb (4SC) and 10-meter walk/run (10MWR), and the 6MWD were all available. Primary analyses were based on pooling data across these six sources, as they had the largest number of prognostic factors in common and available for adjustment.

In the North Star UK database, rise from supine and 10MWR were available, while in CCHMC, 4SC was available along with timed sit to stand and 30 ft walk/run. In the DMD Italian Group, 10MWR and 6MWD were available, but height, and steroid type were not available. Given these differences, three sets of sensitivity analyses were done, separately adding data from: 1) North Star UK, 2) CCHMC, and 3) DMD Italian Group, to the six data sources used in the primary analyses. In each of these sensitivity analyses, the set of prognostic factors used for adjustment was based on what was commonly available across all the included sources (**Table A.4**).

Dystrophin genotypes have been associated with differences in ages at loss of ambulation, long-term changes in function, and other clinical milestones in DMD [27,31,42]. The present study did not include genotypic prognostic factors as not all patients and data sources had genotype

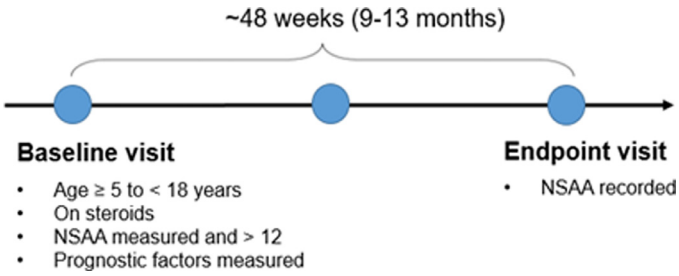


Fig. 1. Study design: schematic for included intervals of follow-up. NSAA, North Star Ambulatory Assessment.

information available. The prognostic value of dystrophin genotypes for 48-week changes in NSAA was also expected to be relatively small and is the subject of a separate study being conducted within cTAP.

### 2.4. Sample selection

Included patients were required to meet each of the following criteria during an interval of follow-up of approximately 48 weeks in length: (1) age ≥ 5 years and < 18 years at the first visit in the interval (referred to as the baseline visit); (2) receiving either prednisone or deflazacort at the baseline visit, (3) baseline NSAA total score > 12; (4) NSAA total score available at the endpoint visit of the interval; and (5) prognostic factors listed above were non-missing at the baseline visit (if available at all in the data source) (**Fig. 1**). Criterion (2) above could not be applied exactly as specified above in the DMD Italian Group data due to lack of availability of data on whether prednisone or deflazacort was the steroid being used. However, data on use of steroids (without mention of steroid type) was available and used to restrict the sample to patients known to be daily or intermittent steroid users.

Within the RWD/NHD sources some patients had multiple ~48-week intervals meeting the above criteria; in such cases, all non-overlapping intervals were included in the analyses. The endpoint visit for one interval was allowed to serve as the baseline visit for the subsequent interval, but further overlap was disallowed.

### 2.5. Statistical analysis

Patient characteristics and ΔNSAA were summarized across patient intervals and by data sources. Statistical comparisons utilized generalized estimating equations (GEEs) with an exchangeable covariance structure to account for use of multiple intervals from individual patients [41].

Mean ΔNSAA was compared between the clinical trial placebo arms and the RWD/NHD, both with and without adjustment for baseline prognostic factors. As adjustment for prognostic factors can be done by applying different statistical methods with different pros and cons, we used two of the most frequently used adjustment methods in this analysis, multivariable regression and propensity score matching [43]. Multivariable regression is the most commonly used method to adjust for baseline differences in observational data and required including the available prognostic factors as covariates in a regression model relating ΔNSAA to data source type. Propensity score matching is an alternative adjustment approach that involved summarizing available prognostic factors into a single value (the propensity score), pairing trial placebo and RWD/NHD patients who had very similar propensity scores, and then comparing differences between trial placebo and RWD/NHD groups in this matched sample. Differences in prognostic factors between groups are eliminated or reduced in this matched sample, thereby allowing for an adjusted estimate of differences between RWD/NHD and trial placebo arms.

#### 2.5.1. Multivariable regression

In the primary analyses, three multivariable linear regression models were fit: (1) an unadjusted model; (2) a *base* model adjusting only for age, steroid type, and baseline NSAA; and (3) a *full* model, additionally including height, weight, body mass index (BMI), 6MWD, rise from supine, 4SC, and 10MWR, which have been identified as prognostic factors for change in ambulatory function from previous research [20,44]. Estimates and 95% confidence intervals (CI) for the difference in mean ΔNSAA between trial placebo arms and RWD/NHD were obtained for each model. Differences in mean ΔNSAA across individual data sources were also investigated. GEEs were used to account for within-subject correlation.

#### 2.5.2. Propensity score matching

Propensity scores were estimated from a logistic regression model incorporating the same covariates included in the regression analyses described above [45]. The propensity score maps each profile of baseline characteristics to a likelihood of appearing in the placebo arms vs. the RWD/NHD. Patients in RWD/NHD were then matched 1:1 to patients in the pooled trial placebo arms using nearest-neighbor matching on the logit of the propensity score, and requiring matches to be within 20% of this measure's pooled standard deviation [46]. Baseline balance between placebo arms and RWD/NHD was assessed using absolute standardized differences, with values < 0.1 interpreted as acceptable balance [43]. The difference in mean ΔNSAA between trial placebo arms and RWD/NHD was then calculated in the propensity-score matched sample. GEEs were used to account for within-subject correlation.

#### 2.5.3. Sensitivity analyses

To assess whether consistency between trial placebo arms and RWD/NHD sources was sensitive to adjustment of different sets of prognostic factors, data from North Star UK, CCHMC, and DMD Italian Group, which included different sets of baseline prognostic factors, were incorporated in three separate sensitivity analyses. Each sensitivity analysis added one RWD/NHD data source to the set of data sources used in the primary analysis: the first added North Star UK, the

Table 2
Summary statistics for unmatched and matched samples (Primary analysis).

| | Unmatched | | | Matched | | |
|---|---|---|---|---|---|---|
| | Placebo arms $N=235$ intervals (235 patients) | RWD/NHD $N=348$ intervals (202 patients) | Absolute Std. diff. | Placebo arms $N=179$ intervals (179 patients) | RWD/NHD $N=179$ intervals (138 patients) | Absolute Std. diff. |
| Age (years) | 9.1 ± 1.8 | 9.1 ± 2.3 | 0.00 | 9.1 ± 1.8 | 9.2 ± 2.2 | 0.02 |
| Height (cm) | 124.4 ± 9.8 | 121.8 ± 10.8 | 0.26 | 123.6 ± 10.1 | 123.6 ± 10.7 | 0.00 |
| Weight (kg) | 29.4 ± 9.3 | 28.5 ± 9.4 | 0.09 | 29.4 ± 10.0 | 29.0 ± 9.1 | 0.05 |
| BMI (kg/m2) | 18.6 ± 3.9 | 18.8 ± 3.7 | 0.04 | 18.8 ± 4.1 | 18.6 ± 3.6 | 0.07 |
| Deflazacort Use, N (%) | 108 (46.0) | 243 (69.8) | 0.50 | 102 (57.0) | 97 (54.2) | 0.06 |
| NSAA total score | 23.3 ± 5.9 | 25.1 ± 6.0 | 0.30 | 24.1 ± 5.9 | 24.0 ± 6.1 | 0.01 |
| 6MWD (m) | 368.0 ± 59.1 | 384.3 ± 76.4 | 0.24 | 375.5 ± 59.2 | 372.7 ± 68.9 | 0.04 |
| Timed 4 stair climb (velocity) (1/s) | 0.26 ± 0.14 | 0.35 ± 0.17 | 0.60 | 0.28 ± 0.15 | 0.28 ± 0.14 | 0.02 |
| Timed 10-meter walk/run (velocity) (m/s) | 1.75 ± 0.51 | 2.04 ± 0.52 | 0.56 | 1.85 ± 0.51 | 1.86 ± 0.47 | 0.02 |
| Timed rise from supine (velocity) (1/s) | 0.17 ± 0.16 | 0.21 ± 0.11 | 0.29 | 0.18 ± 0.11 | 0.18 ± 0.10 | 0.00 |

6MWD: six-minute walk distance; BMI: body mass index; RWD/NHD: real-world data/natural history data; NSAA: North Star Ambulatory Assessment. **Note:** N represents number of ~48-week intervals. Summary statistics in table are Mean ± SD, unless specified otherwise.

second added CCHMC, and the third added DMD Italian Group. Adjustment factors included in each of the three sensitivity analyses were limited to those available in all the included data sources.

#### 2.5.4. Effects of calendar year

Effects of time period on 48-week changes in NSAA in RWD/NHD sources were assessed by adding year categories as covariates into the primary regression analysis. Year categories were selected ad hoc based on the distribution represented in the RWD/NHD sources (**Table A.5** and **Figure A.1**).

#### 2.5.4. Subgroup analysis

Finally, as several current and upcoming trials in ambulatory DMD are evaluating treatments to be administered earlier in life and enrolling patients from younger ages, we also assessed the consistency of mean ΔNSAA across data sources specifically among younger patients, by repeating the multivariable regression analyses in a subgroup of patients age $\geq 5$ to $< 8$ years.

## 3. Results

### 3.1. Baseline characteristics and outcomes

The primary analysis included 437 patients, contributing a total 583 ~48-week intervals of follow-up for ΔNSAA. The three clinical trial placebo arms contributed 235 patients (235 intervals). The pooled RWD/NHD (Leuven, PRO-DMD-01, and iMDEX) included 202 patients (348 intervals).

Patients in the placebo arms had similar mean ages to patients in the RWD/NHD group at baseline (9.1 vs. 9.1 years, respectively) (**Table 2, left panel**). The median age was 8.8 years (range 5.3 to 14.6 years) in the placebo arms and 8.7 (range 5.0 to 16.4) in the RWD/NHD. However, patients in the placebo arms were less likely to be treated with deflazacort (46.0% vs. 69.8%, respectively), and had substantially poorer performance on 6MWD, timed 4SC, timed 10MWR, and timed rise from supine. The median NSAA total score at baseline was 23.0 (range 13 to 34) in the placebo arms and 26.0 (range 13 to 34) in the RWD/NHD. Placebo arms were also enriched for nonsense mutations (via placebo arms from ataluren trials) and patients amenable to skipping of exon 51 (via the DEMAND III placebo arm).

Overall, the distribution of ΔNSAA had a mean of −3.0 (−3.8 in placebo arms, and −2.5 in RWD/NHD). The median (range) of 48-week changes in NSAA total score in placebo arms and RWD/NHD sources were −3.0 (−16.0 to +5.0) and −1.9 (−22.8 to +9.8), respectively.

### 3.2. Assessment of prognostic factors

In the base model, baseline age, NSAA, and steroid type were all statistically significant predictors of ΔNSAA. However, after adjustment for the additional prognostic factors in the full model, age was no longer statistically significant. In the full model, higher baseline NSAA, greater height and higher BMI were associated with mean declines in NSAA over 48 weeks, while greater weight, deflazacort use (vs. prednisone), better performance on 6MWD and timed function tests, were associated with mean increases in NSAA over 48 weeks (**Table 3**). The full model explained 28% of the variability in ΔNSAA whereas the base model accounted for only 17% of variability in ΔNSAA.

### 3.3. Multivariable regression adjustment

Mean (standard deviation [SD]) of ΔNSAA was −3.8 units (4.6) in the pooled placebo arms and −2.5 units (4.2) in the pooled RWD/NHD, indicating similar means and

Table 3
Prognostic factors for ΔNSAA (multivariable regression model).

| Baseline characteristics | Estimate | SE | P-value | |
|---|---|---|---|---|
| Intercept | 34.95 | 7.39 | <0.0001 | * |
| Placebo (reference: RWD/NHD) | 0.12 | 0.35 | 0.72 | |
| Age (yrs) | −0.18 | 0.11 | 0.12 | |
| Height (cm) | −0.34 | 0.06 | <0.0001 | * |
| Weight (kg) | 0.59 | 0.11 | <0.0001 | * |
| BMI | −0.94 | 0.19 | <0.0001 | * |
| Deflazacort (reference: prednisone) | 1.47 | 0.33 | <0.0001 | * |
| NSAA total | −0.16 | 0.05 | 0.0012 | * |
| 6MWD (m) | 0.01 | 0.00 | 0.00086 | * |
| 4-stair climb (1/s) | 5.34 | 1.51 | 0.00041 | * |
| 10 meter walk/run (m/s) | 1.52 | 0.66 | 0.021 | * |
| Rise from floor (1/s) | 2.24 | 2.68 | 0.40 | |

6MWD: six-minute walk distance; BMI: body mass index; RWD/NHD: real-world data/natural history data; NSAA: North Star Ambulatory Assessment; $^{*}p < 0.05$.

levels of dispersion across groups. This difference in mean ΔNSAA between placebo arms and RWD/NHD sources, though small in magnitude, was statistically significant prior to adjustment (−1.2 units, 95% CI: [−2.0, −0.5]). This difference was attenuated and no longer statistically significant after adjustment for prognostic factors, decreasing to −0.6 units (95% CI: [−1.3, 0.1]) in the base model, and 0.1 units (95% CI: [−0.6, 0.8]) in the full model (**Fig. 2 a).**

Mean ΔNSAA in individual data sources ranged between −2.3 and −4.3 units before adjustment, compared to between −2.7 and −3.4 units in the fully adjusted model (**Fig. 2b**); there were no significant differences between individual data sources after adjustment for prognostic factors in the full model.

### 3.4. Propensity score matching

Propensity score matching resulted in 358 matched patient intervals (179 per group) drawn from 317 patients (179 from trial placebo arms; 138 from RWD/NHD sources). Balance in baseline characteristics between the placebo arm and RWD/NHD patients was improved in the propensity score matched sample (**Table 2, right panel**). All covariate measures had absolute standardized differences below the threshold of 0.1 indicating adequate balance between groups.

In the unmatched sample, the difference in mean ΔNSAA between placebo arms and RWD/NHD sources was −1.2 units (95% CI: [−2.0, −0.5]). In the matched sample, these differences were numerically smaller and not statistically different from zero (0.2 units, 95% CI: [−0.7, 1.0]) (**Fig. 3**).

### 3.5. Sensitivity analyses

Results were generally similar in the sensitivity analyses including North Star UK and CCHMC data, respectively (**Table 4**). In the analysis incorporating North Star UK, the unadjusted difference in mean ΔNSAA of −1.1 units, 95% CI: [−1.7, −0.3]) between placebo arms and RWD/NHD was reduced and no longer statistically significant after adjustment

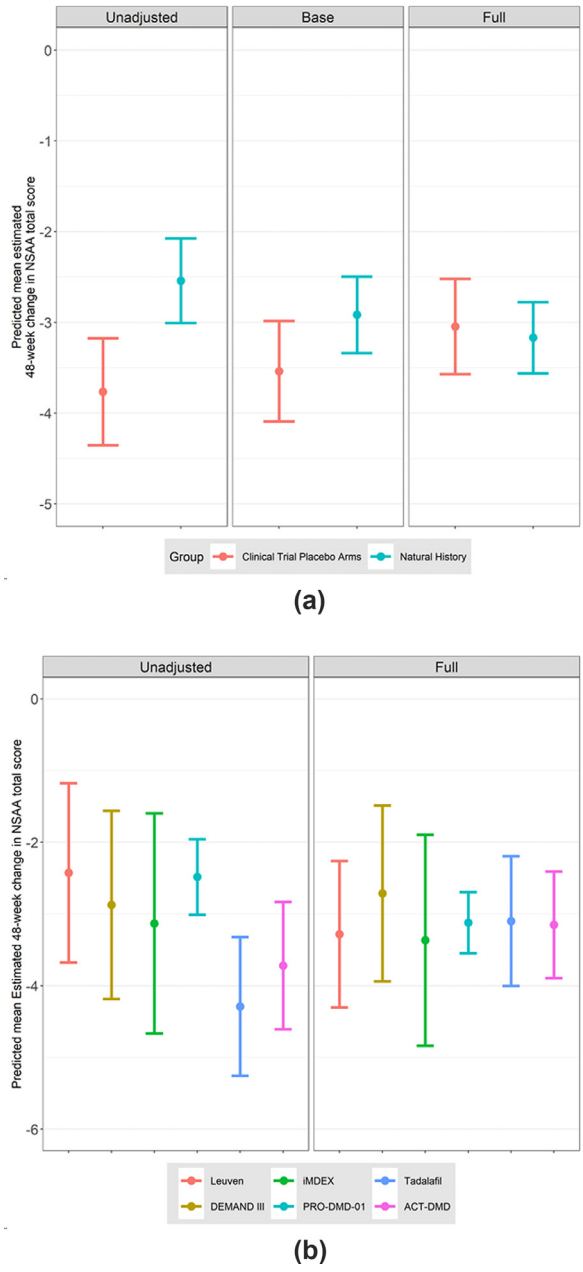


Fig. 2. a. Mean ΔNSAA in RWD/NHD and trial placebo arms in unadjusted and adjusted analyses using multivariable regression. b. Mean ΔNSAA in individual data sources in unadjusted and fully adjusted analyses using multivariable regression. NSAA, North Star Ambulatory Assessment.

in the full model (−0.2 units, 95% CI: [−0.9, 0.4]), and in the propensity-score matched sample (0.2 units, 95% CI: [−0.7, 1.0]). Similarly, in the analysis including CCHMC, a significant difference between placebo and RWD/NHD before adjustment (−1.4 units, 95% CI: [−2.1, −0.8]) was reduced and no longer statistically significant after adjustment in the full model (0.1 units (95% CI: [−0.6, 0.8]), and in the propensity-score matched sample (0.1 units (95% CI: [−0.8, 0.9]). Unlike in the primary analyses in which no significant differences between individual data sources were evident after adjustment for prognostic factors, differences between individual data sources were noted in the two larger

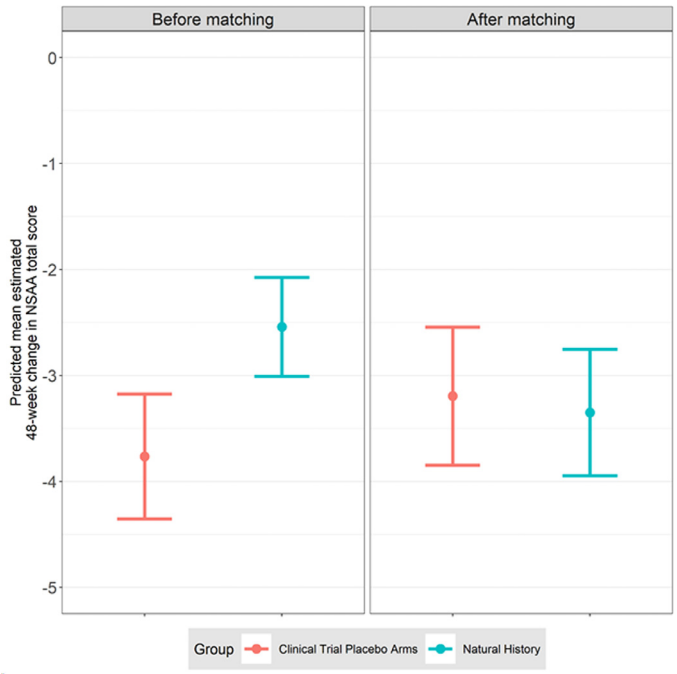


Fig. 3. Mean ΔNSAA in RWD/NHD and trial placebo arms before and after propensity score matching. NSAA, North Star Ambulatory Assessment; NHD, natural history data; RWD, real-world data.

sensitivity analyses samples. In particular, the estimated mean ~48-week decline in NSAA total after regression adjustment was significantly greater in iMDEX than in any of the other data sources (~2 units larger mean decline). There were no differences observed between any of the other pairs of data sources.

Finally, in the sensitivity analysis including DMD Italian Group, the unadjusted difference in mean ΔNSAA of −1.4 units, 95% CI: [−2.1, −0.7]) between placebo arms and RWD/NHD was reduced but remained statistically significant after adjustment in the full model (−0.9 units, 95% CI: [−1.6, −0.3]) and in the propensity-score matched sample (−1.1 units, 95% CI: [−1.9, −0.3]).

### 3.6. Effects of calendar year

Mean changes in 48-week NSAA initially varied across calendar year categories (**Table A.4**, **Figure A.2**). After adjustment for baseline prognostic factors, however, calendar year effects were attenuated and not statistically significant. There was no evidence of a trend towards improving NSAA change outcomes over time in the RWD/NHD after adjustment for baseline status.

### 3.7. Subgroup analysis

The subgroup analysis of younger boys was based on a sample of 214 intervals from 171 patients. In this subgroup of patients, mean age was 7.0 years (median 7.2 years); 54.2% were on deflazacort, and mean (SD) of NSAA total and 6MWD were 25.8 (5.3) and 385.2 (69.6), respectively. Mean (SD) of ΔNSAA was −2.4 units (4.4) from 71 intervals in the pooled placebo arms group, and −1.1 units (4.3) from 143 intervals in the pooled RWD/NHD group. As in the full analyses, differences were significant prior to adjustment (−1.4 units, 95% CI: [−2.7 −0.2]), but attenuated and no longer statistically significant after adjustment for prognostic factors (0.6 units, 95% CI: [−0.5, 1.8]).

Table 4
Sensitivity analyses.

| | Primary analysis | Sensitivity analysis adding North Star UK | Sensitivity analysis adding CCHMC | Sensitivity analysis adding DMD Italian Group |
|---|---|---|---|---|
| **Multivariable regression analyses** | | | | |
| Number of ~48-week intervals (patients) | 583 (437) | 786 (569) | 1018 (635) | 745 (509) |
| Unadjusted | −1.2 (−2.0, −0.5)* | −1.0 (−1.7, −0.3)* | −1.4 (−2.1, −0.8)* | −1.4 (−2.1, −0.7)* |
| Full model | 0.1 (−0.6, 0.8) | −0.2 (−0.9, 0.4) | 0.1 (−0.6, 0.8) | −0.9 (−1.6, −0.3)* |
| **Propensity score matching analyses** | | | | |
| Number of ~48-week intervals (patients) | 358 (317) | 450 (408) | 406 (367) | 466 (411) |
| Before matching | −1.2 (−2.0, −0.5)* | −1.1 (−1.7, −0.3)* | −1.4 (−2.1, −0.7)* | −1.4 (−2.1, −0.7)* |
| After matching | 0.2 (−0.7, 1.0) | 0.2 (−0.7, 1.0) | 0.1 (−0.8, 0.9) | −1.1 (−1.9, −0.3)* |

4SC, 4-stair climb; 6MWD, six-minute walk distance; 10MWR, 10-meter walk run; BMI, body mass index; CCHMC, Cincinnati Children's Hospital Medical Center; CI, confidence interval; DMD, Duchenne muscular dystrophy; NSAA, North Star Ambulatory Assessment.
**Notes:** Table shows difference in mean ΔNSAA (95% CI) between RWD/NHD and trial placebo arms across different analyses. Primary analysis was based on data from the tadalafil DMD trial placebo arm, ACT DMD placebo arm, DEMAND III placebo arm, PRO-DMD-01 study, Leuven database and the iMDEX study. Each sensitivity analysis added the one mentioned data source to the core set of data sources used in the primary analysis. Primary analysis adjusted for age, NSAA total, deflazacort use, height, weight, BMI, 6MWD, and 4SC, 10MWR and rise from supine velocities. Sensitivity analysis including North Star UK adjusted for age, NSAA total, deflazacort use, height, weight, BMI, and 10MWR and rise from supine velocities. Sensitivity analysis including CCHMC adjusted for age, NSAA total, deflazacort use, height, weight, BMI and 4SC velocity. Sensitivity analysis including DMD Italian Group adjusted for age, NSAA total, height, 6MWD, and 10MWR and rise from supine velocities.
* $p<0.05$.

## 4. Discussion

The present study is one of a series assessing the suitability of using NHD and RWD in drug evaluation in DMD; an earlier study focused on 6MWD [37] whereas the present study focused on NSAA as the outcome measure. A primary impetus for these studies was the concern that patients in clinical trials might demonstrate better functional outcomes than in RWD/NHD settings, perhaps due to differences in motivation or other factors, which would bias comparisons to external controls. We found no evidence of such bias in average NSAA outcomes between placebo and RWD/NHD groups in the present study. Rather, we observed a striking similarity in average 48-week changes in NSAA across these groups, especially after adjustment for known baseline prognostic factors. Differences in mean ΔNSAA across data sources were small in magnitude at 1–2 units, which is smaller than minimal detectable change for this measure [47]. Furthermore, due to the large sample sizes available in the present study, the plausible magnitude of differences in mean ΔNSAA between the studied placebo arms and RWD/NHD sources was limited to ± 1 units, as indicated by tight 95% confidence limits. Consistency across placebo arms and RWD/NHD was also observed among younger boys aged 5–8 years. Taken together, these studies of 6MWD [37] and NSAA further support the suitability of using external controls from RWD/NHD, in pre-specified and well-designed studies, for drug evaluations in ambulatory DMD.

Many factors differ between the placebo arms and RWD/NHD sources evaluated in the present study and, in the absence of our findings, would raise unaddressed concerns for use of external controls. Our findings of consistency in mean ΔNSAA across data sources indicates that the net impacts of all of these confounding factors are small. The generalizability of this finding to new clinical trials, other RWD/NHD sources, boys younger than 5 years, and longer-term follow-up needs to be evaluated on a case-by-case basis. Below, we further discuss a number of the potential confounding factors that could have led to differences in ΔNSAA across data sources, the limitations of our study, and overall implications for clinical trial design and analysis in DMD.

### *4.1. Consistency across clinical trial and usual care settings*

We observed consistency in NSAA outcomes despite several known or potential differences between the RWD/NHD sources and trial settings. Compared to the assessment of NSAA in RWD/NHD, patients and assessors in the blinded placebo arms were aware of the possibility that the patient was receiving active study therapy. The prospect of active therapy could have been hypothesized to influence patient motivation or hope for improvement, and to thereby impact NSAA performance. External cues from caregivers and assessors could have also differed between settings, and impacted the NSAA performance or assessment. Inclusion criteria also differed; in particular the clinical trials included baseline thresholds for 6MWD, which was not measured in all RWD/NHD sources, and was not used for selection of our study sample. Despite these specific differences, and other differences that could be hypothesized, we detected no evidence of aggregate, net bias in mean NSAA outcomes between placebo arms and RWD/NHD.

### *4.2. Consistency across time periods, geographies and clinics*

Our studied data sources differed in terms of time periods covered, geographic representation and specific clinics, which in turn might be associated with different care patterns, including steroid treatment, and outcome assessment practices. We observed consistency in mean ΔNSAA outcomes despite these differences in the primary analyses. Investigation of time period effects in the RWD/NHD revealed no meaningful trends over time in NSAA outcomes after accounting for baseline status. In addition, our findings were reassuringly similar across sensitivity analyses incorporating additional data sources. Analyses incorporating the North Star UK, CCHMC, and DMD Italian Group databases, in particular, added patients from two large RWD sources from different countries, with notable differences in care setting (a single center at CCHMC and multiple centers throughout the UK in North Star UK and throughout Italy in the Italian Group), steroid prescribing (predominantly prednisone in North Star UK, compared to predominantly deflazacort in CCHMC), and different functional measures available for baseline adjustment.

Average NSAA outcomes were highly consistent, after adjustment for baseline prognostic factors, across the studied care centers and time periods primarily representing the years 2006–2016. Consequently, external control groups in ambulatory DMD may not be biased by differences in geography or time period within these ranges. However, the longevity of these data sources for use as external controls into the future will need to be continually evaluated on a case-by-case basis. External controls should be drawn from time periods and geographies that are comparable to those of the treated patients, and ideally from parallel groups, to ensure that differences in standards of care do not confound outcomes.

### *4.3. Consistency across different baseline adjustment methods*

Our findings were also consistent across two different statistical methods for baseline adjustment: multivariable regression and propensity score matching. This suggests that, for avoidance of bias, adjustment for an adequate set of baseline prognostic factors is more important than the choice of statistical method used for adjustment in this case. Multivariable regression did provide more precise estimates (smaller standard errors and narrower CIs) than propensity score matching. This was due to the fact that including adjustment factors in a model helps to explain some of the variation in the outcome [29,30] and, less so, to the

incorporation of more of the data (since there was no need to exclude unmatched subjects).

### 4.4. Prognostic factors for change in NSAA

The baseline factors identified as prognostic for change in NSAA were highly consistent with those previously identified and discussed as prognostic for changes in 6MWD [12]. Multiple baseline functional measures were highly prognostic, as were steroid type, height, weight, and BMI. Age did not add significant prognostic value for 48-week changes in NSAA total score, beyond that provided by these other factors, but was readily available and was always included for adjustment in the present study. Interpretations and implications of these prognostic factors, and consideration of their correlation with each other and the fact that BMI is itself based on height and weight, has been discussed previously [14]. Knowledge of important prognostic factors is critical for comparisons to external controls. Our findings for NSAA, in combination with earlier results for 6MWD and timed 4SC [20,44], indicate that combinations of baseline functional measures are strongly prognostic, and should be used for matching or adjustment in comparisons to external controls in DMD. While we did not attempt to identify an optimal set of baseline factors to adjust for, we observed that adjustment for steroid type, height, weight, BMI, and multiple functional characteristics – as in our primary analysis, which adjusted for 6MWD, NSAA, and velocities for rise from supine, 4SC, and 10MWR – provided the greatest reduction in the already small differences in mean ΔNSAA across data sources. In the sensitivity analyses incorporating DMD Italian Group, in which, unlike the primary and other sensitivity analyses, adjustment was not possible for steroid type, weight, and BMI, small differences remained between placebo arms and RWD/NHD sources, indicating the incremental importance of adjusting for these factors. Potential impacts of differences in steroid dosing, regimen and age of initiation were not accounted for as these data were not systematically available across the data sources analyzed here. When available, the assessment of and adjustment for baseline differences in these factors may also be relevant for comparisons versus external controls.

Differences in dystrophin genotypes are known to impact long-term outcomes in DMD [31]. While dystrophin genotypes might not be expected to have as large effects on 48-week changes in NSAA as they do on longer-term outcomes and milestones, their prognostic associations warrant further study, which is separately underway within our collaboration [48].

### 4.5. Generalizability to other RWD/NHD sources

The data sources included in this study shared important similarities that likely contributed to the consistency in mean ΔNSAA outcomes between RWD/NHD and placebo arms. Firstly, the RWD/NHD databases represent care centers with extensive research and DMD clinical trial experience, with NSAA assessment typically conducted by physicians, physical therapists, or clinical evaluators who are trained and experienced in administering the NSAA and other functional assessments in clinical trial settings. Our findings may or may not generalize to care centers that have less experience with NSAA assessments. In addition, the time periods represented by the RWD/NHD sources studied here precede the approvals of multiple targeted therapies for DMD. As more therapies are approved, differences in treatment use could introduce further variation in outcomes across care centers and data sources. The suitability of additional follow-up data from these RWD/NHD sources, or of data from other RWD/NHD sources, will need to be evaluated case-by-case for provision of external controls.

### 4.6. Limitations and considerations for application to external controls

There are several issues not fully addressed in the present study that will be relevant for formal comparison of RWD/NHD external controls and specific treatment groups. Firstly, there are additional potential prognostic factors that could be evaluated for baseline adjustment. While our analyses focused on steroid users and adjusted for steroid type where possible, data on steroid regimen and age of initiation were not consistently available in these RWD/NHD sources, would likely vary across time and care centers, and could potentially impact patients' outcomes. Data on lower extremity contractures may affect functional test performance but were not available for adjustment in this analysis.

Secondly, data quality, in terms of collection processes, standardization, recording, verification, and other factors would need to be assessed across data sources used to form an external control group for regulatory evaluation. Mechanisms that lead to missing data, and conventions for handling missing data, on functional tests such as the NSAA can vary across data sources. Going forward, harmonization and standardized documentation of reasons for NSAA item non-assessment would be valuable to address this limitation and maximize the research value of NSAA data collected from all assessments. Thirdly, it should be noted that comparisons between RWD/NHD and single-arm or uncontrolled studies would require an additional layer of caution, as patients included in single-arm trials are certain that they are receiving active therapy, whereas patients in the placebo arms included in the present study had only a probability of receiving blinded active therapy.

An alternative to replacing a placebo controlled trial design is to augment a smaller placebo arm with RWD/NHD [39–41], an approach which has been explored in DMD [49]. This approach benefits from the incorporation of some randomized placebo treated patients, but may exacerbate the challenge of achieving adequate randomization to be representative of the drug treated arm(s). Indeed, an augmented analysis can, appropriately, have worse power than a simple analysis of the randomized data if outcomes are inconsistent between randomized placebo and the external controls that have been pre-specified for augmentation [32]. Statistical methods for

using RWD/NHD in DMD drug evaluation, including placebo augmentation [32–34] and comparison of observed outcomes for treated patients with predictions of their untreated outcomes given baseline or pre-baseline prognostic factors [50], warrant further development and evaluation for statistical performance in DMD. Importantly, all of these approaches can build on the foundation provided by the present findings, since all methods of incorporating external controls have a fundamental reliance on consistency in outcome measures across data sources and understanding of important prognostic factors.

Finally, while we have shown consistency of 48-week changes in NSAA and 6MWD between RWD/NHD and placebo arms, the viability of using RWD/NHD data as longer-term external comparator groups for longer-term trials or for trial extension periods needs to be assessed further.

### *4.7. Complementary evidence for consistency of NSAA and 6MWD*

It is noteworthy that a high degree of consistency between RWD/NHD and placebo arms has now been demonstrated for 48-week changes in both the NSAA, in the present study, and the 6MWD [37]. These commonly used functional endpoints differ in how they are assessed, and in their potential sources of bias with the NSAA being a primarily clinician-reported outcome and 6MWD being a performance outcome [51]. For 6MWD, even with standardized test settings, concerns had centered on the risk that motivation or effort level on this rather long and burdensome test could vary across settings and lead to differences in performance. The NSAA, on the other hand, requires patients to perform a less burdensome set of activities, but may be considered more sensitive to evaluators' judgements – especially for differentiating between a score of a '1' or '2' for a particular task. The importance of evaluator training for the reliability of the NSAA was highlighted early in its development [25,26], and led to sustained efforts to standardize training across centers performing this test. Across the data sources included in this study, which all incorporated standardized training for the NSAA, the consistency in mean ΔNSAA is notable amid the variation in geographies and time periods represented.

### *4.8. Collaboration*

Our goal of understanding the consistency of NSAA between RWD/NHD and placebo arm settings was accomplished by collaboratively analyzing a broad collection of data sources. The sharing of clinical data across multiple institutions, registries, and geographies was a challenge, but one worth overcoming for this important research goal. Collaborating through cTAP simplified and accelerated this process, and highlights the importance of data collection, data sharing, and collaboration for DMD drug development. Careful, evidence-based use of external controls can potentially reduce the number of patients that need to receive placebo while advancing the development of effective therapies in DMD. We conclude that the high degree of consistency in NSAA outcomes observed in the present study, the demonstrated performance of adjustment for known prognostic factors, and previously observed consistency for 6MWD outcomes, together provide a strong foundation for the incorporation of external controls into the evaluation of drug effects on these outcomes in DMD. We emphasize that no externally controlled study can match the rigor of a well-powered, randomized, placebo-controlled study – and that any use of external controls needs to be evaluated on a case-by-case basis and in light of the findings presented here.

## Previous presentations

World Muscle Society 2019, Action Duchenne 2019, The Professional Society for Health Economics and Outcomes Research Europe (ISPOR EU) 2019

## Author contributions

Conceptualization: FM, JS, GS, SJW, NG, BW, EM
Data curation: FM, NG, BW, CT, EM, AM, IJdG, EHN, LS, VS, CM
Formal analysis: JS, GS, ND, HW, JM, ZY
Writing - original draft: JS, GS, ND, SJW
Writing - review & editing: All authors

## Declaration of Competing Interest

Francesco Muntoni is a member of the Rare Disease Scientific Advisory Group for Pfizer and of Dyne Therapeutics SAB, and has participated to SAB meetings for PTC, Sarepta, Santhera, Wave Therapeutics. UCL and Great Ormond Street Hospital are recipient of grants from Pfizer, Italfarmaco, Wave, Santhera, Sarepta regarding clinical trials. James Signorovitch co-founded the collaborative Trajectory Analysis Project (cTAP) and is an employee of Analysis Group, Inc., a consulting firm that received funding from the membership of cTAP to conduct this study. Gautam Sajeev, Nicolae Done, Hallee Wong, Jackson Moss, and Zhiwen Yao are employees of Analysis Group, Inc., a consulting firm that received funding from the membership of cTAP to conduct this study. Nathalie Goemans has received compensation for consultancy services from Eli Lilly, Italfarmaco, PTC Therapeutics, BioMarin Pharmaceutical, Pfizer, Avidity, Daiichi Sankyo, Wave, Santhera and has served as site investigator for GlaxoSmithKline, Prosensa, BioMarin Pharmaceutical, Italfarmaco, Eli Lilly, Wave, and Sarepta. Brenda Wong has participated in advisory committee meetings for Prosensa and Biomarin and has received compensation for consultancy services for Gilead Sciences, Pfizer and GSK. Cuixia Tian has served as the site investigator for trials sponsored by PTC Therapeutics, Eli Lilly, GSK, Prosensa/Biomarin, Bristol Myers Squibb, Roche, Pfizer. Santhera, Sarepta, Fibrogen, Capricor, Pfizer, Avexis, and Catabasis. Eugenio Mercuri has served on clinical steering

committees and/or as a consultant for Eli Lilly, Italfarmaco, PTC Therapeutics, Sarepta, Santhera, and Pfizer; has served as PI for GlaxoSmithKline, Prosensa, BioMarin Pharmaceutical, Italfarmaco, Roche, PTC, Pfizer, Sarepta, Santhera, Wave, NS and Eli Lilly. Susan J. Ward co-founded and manages the collaborative Trajectory Analysis Project and has received funding from the membership of cTAP to facilitate this study. Adnan Manzur has no disclosures. Laurent Servais is member of the SAB or has performed consultancy for Sarepta, Dynacure, Santhera, Avexis, Biogen, Cytokinetics and Roche, Audentes Therapeutics and Affinia Therapeutics; Laurent Servais has given lectures and has served as a consultant for Roche, Biogen, Avexis, and Cytokinetics. Laurent Servais is the project leader of the newborn screening in Southern Belgium funded by Avexis, Roche, and Biogen. Erik H. Niks is a member of the European Reference Network for Rare Neuromuscular Diseases (ERN EURO-NMD). Erik H. Niks reports grants from Duchenne Parent Project, ZonMW and AFM, consultancies for BioMarin and Summit, and worked as local investigator of clinical trials of BioMarin, GSK, Lilly, Santhera, Givinostat, and Roche outside the submitted work. E.H.N. reports ad hoc consultancies for WAVE, Santhera, Regenxbio, and PTC, and he worked as investigator of clinical trials of Italfarmaco, NS Pharma, Reveragen, Roche, WAVE, and Sarepta outside the submitted work. Volker Straub has participated in advisory boards for Audentes Therapeutics, Biogen, Exonics Therapeutics, Italfarmaco S.p.A., Roche, Sanofi Genzyme, Sarepta Therapeutics, Summit Therapeutics, UCB, and Wave Therapeutics. He has research collaborations with Ultragenyx and Sanofi Genzyme. Imelda JM de Groot has no disclosures. Craig McDonald has served as a consultant for PTC Therapeutics, BioMarin Pharmaceutical, Sarepta Therapeutics, Eli Lilly, Pfizer Inc, Santhera Pharmaceuticals, Cardero Therapeutics, Inc, Catabasis Pharmaceuticals, Capricor Therapeutics, Astellas Pharma (Mitobridge), and FibroGen, Inc; serves on external advisory boards related to DMD for PTC Therapeutics, Sarepta Therapeutics, Santhera Pharmaceuticals, and Capricor Therapeutics; and reports grants from US Department of Education/National Institute on Disability and Rehabilitation Research, the National Institute on Disability, Independent Living, and Rehabilitation Research, US NIH/National Institute of Arthritis and Musculoskeletal and Skin Diseases, NIH/National Institute of Neurologic Disorders and Stroke, US Department of Defense, and Parent Project Muscular Dystrophy US.

## Acknowledgements

The authors are grateful to patients for participating in the clinical assessments and agreeing to make their data available for research. The authors also thank Eli Lilly for contributing placebo arm data from the tadalafil DMD trial. Gloria DeWalt, PhD, an employee of Analysis Group, Inc., assisted with the development of this manuscript. The authors would like to acknowledge investigators and staff from each of the RWD/NHD sources included in this study. The iMDEX study is funded by the Association Française contre les Myopathies (AFM), while the UK North Star is funded by Muscular Dystrophy UK. Francesco Muntoni is partially supported by the National Institute for Health Research (NIHR) Biomedical Research centre at Great Ormond Street Hospital for Children NHS Foundation Trust and University College London.

*Universitaire Ziekenhuizen Leuven*

M. Van den Hauwe

*PRO-DMD-01 study investigators*

N. Goemans, N. Deconicnk, M. Tulinius, E. Niks, I. de Groot, K. Flanigan, E. Henricson, B. Wong, M. Bernadete Dutra de Resende, G. Vita, U. Schara, J. Kirschner, E. Mercuri, H. Topaloglu, D. Monges, M. Soledad, C. Cances

*iMDEX*

V. Selby, D. Ridout, V. Decostre, A. Mayhew, M. Guglieri, M. Van der Holst, M. Jansen, J.J.G.M. Verschuuren, I.J.M. de Groot, E.H. Niks, L. Servais, V. Straub, J.Y. Hogrel, F. Muntoni

*North Star UK*

F. Muntoni, A.Y. Manzur, S. Robb, R. Quinlivan, A. Sarkozy, P. Munot, M. Main, L.E. Abbot, H. Patel, S. Samsuddin, V. Ayyar-Gupta, K. Bushby, V. Straub, M. Guglieri, C. Bertolli, A. Mayhew, R. Muni-Lofra, M. James, D. Moat, J. Sodhi, H. Roper, D. Parasuraman, H. McMurchie, R.M. Rabb, A. Childs, K. Pysden, L. Pallant, S. Spinty, G. Peachey, R. Madhu, A.J. Shillington, E. Wraige, H. Jungbluth, V. Gowda, J. Sheehan, R. Spahr, I. Hughes, E. Bateman, C. Cammiss, T. Willis, L. Groves, N. Emery, P. Baxter, M.T. Ong, N. Goulborne, M. Senior, E. Scott, L. Hartley, L.B. Parsons, A. Majumdar, K. Vijaykumar, F.F. Mason, L. Jenkins, B. Toms, C.H. Frimpong-Ansah, K. Naismith, J. Dalgleish, A. Keddie, I. Horrocks, M. Di Marco, J. Dunne, G.C.S. Chow, A. Miah, C. de Goede, A. Selley, N. Thomas, M. Illingworth, M. Geary, J. Palmer, C.P. White, K. Greenfield, S. Tiraputhi, S. MacAuley, N. Hussain, H. Robbins, M. Iqbal, G. Ambegaonkar, D. Krishnakumar, C. Ward, J. Taylor, A. O'Hara, J. Tewnion, S.R. Chandratre, S. Ramdas, M. White, H. Ramjattan, A. Baxter, J. Yirrel

*CCHMC*

Neuromuscular providers (B. Wong, C. Tian, I. Rybalsky, K. Shellenbarger) and physical therapists (A. McCormick, M. McGuire, K. Bonnarrigo, A. Fowler, M. Kiefer) and research support (J. Bange, S. Hu)

*DMD Italian Group investigators*

E. Mercuri, M. Pane, E. Mazzone, S. Messina, G.L. Vita, M.P. Sormani, A .D'Amico, E. Bertini, A. Berardinelli, F.

Magri, G. Pietro Comi, G. Baranello, T. Mongini, A. Pini, R. Battini, E. Pegoraro, C. Bruno, L. Politano, S. Previtali

## Funding

This study was conducted within the collaborative Trajectory Analysis Project (cTAP), a pre-competitive coalition of academic clinicians, drug developers, and patient foundations formed in 2015 to overcome the challenges of high variation in clinical trials in DMD. cTAP has received sponsorship from Astellas (Mitobridge), BioMarin Pharmaceutical, Bristol Meyers Squibb, Catabasis, Edgewise Therapeutics, FibroGen, Italfarmaco SpA, Marathon Pharmaceuticals, NS Pharma, Pfizer, PTC Therapeutics, Roche, Sarepta Therapeutics, Shire, Solid Biosciences, Summit Therapeutics, Vertex Pharmaceuticals, Parent Project Muscular Dystrophy, Charley's Fund, and CureDuchenne, a founding patient advocacy partner and provider of initial seed funding to cTAP. Physical function testing at Universitaire Ziekenhuizen Leuven was funded by Fonds Spierzieke Kinderen. The PRO-DMD-01 study was sponsored by BioMarin Pharmaceuticals and data were provided to cTAP by CureDuchenne.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.nmd.2022.02.009.

## References

[1] Wu J, Wang C, Toh S, Pisa FE, Bauer L. Use of real-world evidence in regulatory decisions for rare diseases in the United States-Current status and future directions. Pharmacoepidemiol Drug Saf 2020;29:1213–18. doi:10.1002/pds.4962.

[2] U.S. Food and Drug Administration FDA approves first treatment for a form of Batten disease; 2017. https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm555613.htm [accessed 5 October 2018].

[3] U.S. Food and Drug Administration Drug approval package: Myozyme (Alglucosidase Alfa); 2006. https://www.accessdata.fda.gov/drugsatfda_docs/nda/2006/125141s000_MyozymeTOC.cfm [accessed 5 October 2018].

[4] Jahanshahi M, Gregg K, Davis G, Ndu A, Miller V, Vockley J, et al. The use of external controls in FDA regulatory decision making. Ther Innov Regul Sci 2021;55:1019–35. doi:10.1007/s43441-021-00302-y.

[5] Cyna AM, Costi D, Middleton P. Viewpoint: randomised controlled trials using invasive 'placebo' controls are unethical and should be excluded from Cochrane Reviews. Cochrane Database Syst Rev 2011:ED000029. doi:10.1002/14651858.ED000029.

[6] Margraf J, Ehlers A, Roth WT, Clark DB, Sheikh J, Agras WS, et al. How "blind" are double-blind studies? J Consult Clin Psychol 1991;59:184–7. doi:10.1037//0022-006x.59.1.184.

[7] Therapeutics Sarepta. Sarepta therapeutics announces top-line results for Part 1 of study 102 evaluating SRP-9001, its investigational gene therapy for the treatment of duchenne muscular dystrophy; 2021. https://investorrelations.sarepta.com/news-releases/news-release-details/sarepta-therapeutics-announces-top-line-results-part-1-study-102 [accessed 15 February 2022].

[8] Peay HL, Biesecker BB, Wilfond BS, Jarecki J, Umstead KL, Escolar DM, et al. Barriers and facilitators to clinical trial participation among parents of children with pediatric neuromuscular disorders. Clin Trials 2018;15:139–48. doi:10.1177/1740774517751118.

[9] Merlini L, Sabatelli P. Improving clinical trial design for Duchenne muscular dystrophy. BMC Neurol 2015;15:153. doi:10.1186/s12883-015-0408-z.

[10] Mercuri E, Muntoni F, Osorio AN, Tulinius M, Buccella F, Morgenroth LP, et al. Safety and effectiveness of ataluren: comparison of results from the STRIDE Registry and CINRG DMD natural history study. J Comp Eff Res 2020;9:341–60. doi:10.2217/cer-2019-0171.

[11] Clemens PR, Rao VK, Connolly AM, Harper AD, Mah JK, Smith EC, et al. Safety, tolerability, and efficacy of viltolarsen in boys with duchenne muscular dystrophy amenable to exon 53 skipping: a phase 2 randomized clinical trial. JAMA Neurol 2020;77:982–91. doi:10.1001/jamaneurol.2020.1264.

[12] Mendell JR, Goemans N, Lowes LP, Alfano LN, Berry K, Shao J, et al. Longitudinal effect of eteplirsen versus historical control on ambulation in Duchenne muscular dystrophy. Ann Neurol 2016;79:257–71. doi:10.1002/ana.24555.

[13] Jarow JP, LaVange L, Woodcock J. Multidimensional evidence generation and FDA regulatory decision making: defining and using "real-world" data. JAMA 2017;318:703–4. doi:10.1001/jama.2017.9991.

[14] Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-world evidence - what is it and what can it tell us? N Engl J Med 2016;375:2293–7. doi:10.1056/NEJMsb1609216.

[15] U.S. Food and Drug Administration Framework for FDA's Real-World Evidence program; 2018. https://www.fda.gov/media/120060/download [accessed 27 September 2019].

[16] U.S. Food and Drug Administration Submitting documents using real-world data and real-world evidence to FDA for drugs and biologics. Draft guidance for industry; 2019. https://www.fda.gov/media/124795/download [accessed 15 February 2022].

[17] Eichler HG, Bloechl-Daum B, Broich K, Kyrle PA, Oderkirk J, Rasi G, et al. Data rich, information poor: can we use electronic health records to create a learning healthcare system for pharmaceuticals? Clin Pharmacol Ther 2019;105(4):912–22 Apr. doi:10.1002/cpt.1226.

[18] Eichler H. Real world evidence (RWE) – an introduction; how is it relevant for the medicines regulatory system?; 2018. https://www.ema.europa.eu/documents/presentation/presentation-real-world-evidence-rwe-introduction-how-it-relevant-medicines-regulatory-system-emas_en.pdf [accessed 15 February 2022].

[19] Moseley J. Regulatory perspective on real world wvidence (RWE) in scientific advice; 2018. https://www.ema.europa.eu/en/documents/presentation/presentation-regulatory-perspective-real-world-evidence-rwe-scientific-advice-emas-pcwp-hcpwp-joint_en.pdf [accessed 15 February 2022].

[20] Goemans N, Vanden Hauwe M, Signorovitch J, Swallow E, Song J. Collaborative trajectory analysis P. individualized prediction of changes in 6-minute walk distance for patients with duchenne muscular dystrophy. PLoS ONE 2016;11:e0164684. doi:10.1371/journal.pone.0164684.

[21] U.S. Food and Drug Administration Duchenne muscular dystrophy and related dystrophinopathies: developing drugs for treatment guidance for industry; 2018. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/duchenne-muscular-dystrophy-and-related-dystrophinopathies-developing-drugs-treatment-guidance [accessed 8 October 2020].

[22] U.S. Food and Drug Administration FDA briefing document. Peripheral and central nervous system drugs advisory committee Meeting; 2016. April 25NDA 206488 Eteplirsen https://www.fda.gov/downloads/advisorycommittees/committeesmeetingmaterials/drugs/peripheralandcentralnervoussystemdrugsadvisorycommittee/ucm497063.pdf 2016; [accessed 8 October 2020].

[23] ICH Expert Working Group Choice of control group and related issues in clinical trials E10; 2001. https://www.ema.europa.eu/en/ich-e10-choice-control-group-clinical-trials [accessed 8 October 2020].

[24] Pocock SJ. The combination of randomized and historical controls

in clinical trials. J Chronic Dis 1976;29:175–88. doi:10.1016/0021-9681(76)90044-8.
[25] Scott E, Eagle M, Mayhew A, Freeman J, Main M, Sheehan J, et al. Development of a functional assessment scale for ambulatory boys with Duchenne muscular dystrophy. Physiother Res Int 2012;17:101–9. doi:10.1002/pri.520.
[26] Mazzone ES, Messina S, Vasco G, Main M, Eagle M, D'Amico A, et al. Reliability of the North Star Ambulatory Assessment in a multicentric setting. Neuromuscul Disord 2009;19:458–61. doi:10.1016/j.nmd.2009.06.368.
[27] Ricotti V, Ridout DA, Pane M, Main M, Mayhew A, Mercuri E, et al. The NorthStar Ambulatory Assessment in Duchenne muscular dystrophy: considerations for the design of clinical trials. J Neurol Neurosurg Psychiatry 2016;87:149–55. doi:10.1136/jnnp-2014-309405.
[28] Weiss RB, Vieland VJ, Dunn DM, Kaminoh Y, Flanigan KM. United Dystrophinopathy P. Long-range genomic regulators of THBS1 and LTBP4 modify disease severity in duchenne muscular dystrophy. Ann Neurol 2018;84:234–45. doi:10.1002/ana.25283.
[29] Kahan BC, Jairath V, Dore CJ, Morris TP. The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. Trials 2014;15:139. doi:10.1186/1745-6215-15-139.
[30] Thompson DD, Lingsma HF, Whiteley WN, Murray GD, Steyerberg EW. Covariate adjustment had similar benefits in small and large randomized controlled trials. J Clin Epidemiol 2015;68:1068–75. doi:10.1016/j.jclinepi.2014.11.001.
[31] Bello L, Morgenroth LP, Gordish-Dressman H, Hoffman EP, McDonald CM, Cirak S, et al. DMD genotypes and loss of ambulation in the CINRG Duchenne Natural History Study. Neurology 2016;87:401–9. doi:10.1212/WNL.0000000000002891.
[32] Viele K, Berry S, Neuenschwander B, Amzal B, Chen F, Enas N, et al. Use of historical control data for assessing treatment effects in clinical trials. Pharm Stat 2014;13:41–54. doi:10.1002/pst.1589.
[33] Han B, Zhan J, John Zhong Z, Liu D, Lindborg S. Covariate-adjusted borrowing of historical control data in randomized clinical trials. Pharm Stat 2017;16:296–308. doi:10.1002/pst.1815.
[34] Hobbs BP, Carlin BP, Mandrekar SJ, Sargent DJ. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. Biometrics 2011;67:1047–56. doi:10.1111/j.1541-0420.2011.01564.x.
[35] McDonald CM, Campbell C, Torricelli RE, Finkel RS, Flanigan KM, Goemans N, et al. Ataluren in patients with nonsense mutation Duchenne muscular dystrophy (ACT DMD): a multicentre, randomised, double-blind, placebo-controlled, phase 3 trial. Lancet 2017;390:1489–98 London, England. doi:10.1016/s0140-6736(17)31611-2.
[36] McDonald CM, Henricson EK, Abresch RT, Han JJ, Escolar DM, Florence JM, et al. The cooperative international neuromuscular research group Duchenne natural history study–a longitudinal investigation in the era of glucocorticoid therapy: design of protocol and the methods used. Muscle Nerve 2013;48:32–54. doi:10.1002/mus.23807.
[37] Goemans N, Signorovitch J, Sajeev G, Yao Z, Gordish-Dressman H, McDonald CM, et al. Suitability of external controls for drug evaluation in Duchenne muscular dystrophy. Neurology 2020;95 e1381-e91. doi:10.1212/WNL.0000000000010170.
[38] Victor RG, Sweeney HL, Finkel R, McDonald CM, Byrne B, Eagle M, et al. A phase 3 randomized placebo-controlled trial of tadalafil for Duchenne muscular dystrophy. Neurology 2017;89:1811–20. doi:10.1212/wnl.0000000000004570.
[39] Goemans N, Mercuri E, Belousova E, Komaki H, Dubrovsky A, McDonald CM, et al. A randomized placebo-controlled phase 3 trial of an antisense oligonucleotide, drisapersen, in Duchenne muscular dystrophy. Neuromuscul Disord 2018;28:4–15. doi:10.1016/j.nmd.2017.10.004.
[40] A Prospective Natural History Study of Progression of Subjects With Duchenne Muscular Dystrophy. https://clinicaltrials.gov/ct2/show/NCT01753804 [accessed 15 February 2022 ]
[41] Zeger SL, Liang K-Y, Albert PS. Models for longitudinal data: a generalized estimating equatio1n approach. Biometrics 1988:1049–60.
[42] Zhang S, Qin D, Wu L, Li M, Song L, Wei C, et al. Genotype characterization and delayed loss of ambulation by glucocorticoids in a large cohort of patients with Duchenne muscular dystrophy. Orphanet J Rare Dis 2021;16:188. doi:10.1186/s13023-021-01837-x.
[43] Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res 2011;46:399–424. doi:10.1080/00273171.2011.568786.
[44] Goemans N, Wong B, Van den Hauwe M, Signorovitch J, Sajeev G, Cox D, et al. Prognostic factors for changes in the timed 4-stair climb in patients with Duchenne muscular dystrophy, and implications for measuring drug efficacy: a multi-institutional collaboration. PLoS ONE 2020;15:e0232870. doi:10.1371/journal.pone.0232870.
[45] Rosenbaum P, Rubin D. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat 1985;39:33–8.
[46] Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharm Stat 2011;10:150–61. doi:10.1002/pst.433.
[47] Muntoni F, Signorovitch J, Sajeev G, Done N, Yao Z, Goemans N, et al. Minimal detectable changes in functional measures in duchenne muscular dystrophy: a study of multiple centers, networks and trial arms. In: Musclar dystrophy association virtual clinical & scientific conference; 2021.
[48] Muntoni F, Signorovitch J, Sajeev G, Lane H, Jenkins M, Dieye I, et al. Associations between dystrophin genotype and ambulatory outcomes in DMD, and implications for trials of genotype-targeted therapies. World Muscle Soc 2021.
[49] Lake SL, Quintana MA, Broglio K, Panagoulias J, Berry SM, Panzara MA. Bayesian adaptive design for clinical trials in Duchenne muscular dystrophy. Stat Med 2021. doi:10.1002/sim.9021.
[50] Fouarge E, Monseur A, Boulanger B, Annoussamy M, Seferian AM, De Lucia S, et al. Hierarchical Bayesian modelling of disease progression to inform clinical trial design in centronuclear myopathy. Orphanet J Rare Dis 2021;16:3. doi:10.1186/s13023-020-01663-7.
[51] U.S. Food and Drug Administration Guidance for industry patient-reported outcome measures: use in medical product development to support labeling claims; 2009. https://www.fda.gov/media/77832/download [accessed 8 October 2020].