Nusinersen for treating spinal muscular atrophy: A Single Technology Appraisal Erratum

Produced by	School of Health and Related Research (ScHARR), The University of Sheffield
Authors	Paul Tappenden, Reader, ScHARR, University of Sheffield, Sheffield, UK Jean Hamilton, Research Fellow, ScHARR, University of Sheffield, Sheffield, UK Eva Kaltenthaler, Professor, ScHARR, University of Sheffield, Sheffield, UK Emma Hock, Research Fellow, ScHARR, University of Sheffield, Sheffield, UK Andrew Rawdin, Research Assistant, ScHARR, University of Sheffield, Sheffield, UK Clara Mukuria, Research Fellow, ScHARR, University of Sheffield, Sheffield, UK Mark Clowes, Information Specialist, ScHARR, University of Sheffield, Sheffield, UK Anita Simonds, Professor of Respiratory and Sleep Medicine, Royal Brompton & Harefield NHS Foundation Trust, London, UK Anne-Marie Childs, Consultant Paediatric Neurologist, Leeds Teaching Hospitals, Leeds, UK
Correspondence Author	Paul Tappenden, Reader, ScHARR, University of Sheffield, Sheffield, UK
Date completed	20th June 2018

Table 1: ENDEAR analysis sets (adapted from CS, Table 15)

Analysis	Number of patients	Description
Interim (15 June 2016)	Nusinersen: 51 Sham control: 27	Infants in the ITT set who were assessed at the day 183, 302, or 394 visit and had a time difference of at least 190 days between the date of first dose and the data cut-off date of the interim analysis
Final efficacy set (21 November 2016)	Nusinersen: 73; Sham control: 37	Infants in the ITT set who were assessed at the day 183, 302, or 394 visit and had a time difference of at least 190 days between the date of the first dose and the data cut-off date of the final analysis
Final ITT set (21 November 2016)	Nusinersen: 80; Sham control: 41	All infants who were randomised and received ≥1 dose of study drug

ITT – intention-to-treat

Motor function

Motor function was measured in the ENDEAR study using three measures: Module 2 of the Hammersmith Infant Neurological Examination (HINE-2 - the primary endpoint); the Children's Hospital of Philadelphia Infant Test of Neuromuscular Disorders (CHOP INTEND) and the Compound Muscle Action Potential (CMAP), an electrophysiological technique used to measure nerve function, were both secondary outcomes. Responders were infants with a greater number of motor milestone categories with improvement than worsening[4] (see footnote to Table 9). Motor function outcomes are shown in Table 9.

Table 2: ENDEAR motor function outcomes (adapted from CS, Table 19)

Outcome	Nusinersen	Control	Difference (95% CI); *p*-value
Interim analysis (data cut-off 15 June 2016) (interim analysis set)			
HINE-2 proportion responders	21 (41%)	0 (0%)	41.18 (18.6, 61.20); $p<0.001$
Final analysis (data cut-off 21 November 2016) (efficacy analysis set)			
HINE- 2 proportion responders	37 (51%)	0 (0%)	[redacted] $p<0.0001$
HINE -2 proportion with improvement in total score	49 (67%)	5 (14%)	
HINE -2 proportion with worsening in total score	1 (1%)	8 (22%)	
CHOP INTEND proportion with ≥ 4 point improvement	52 (71%)	1 (3%)	[redacted] $p<0.001$
CHOP INTEND proportion with any improvement	53 (73%)	1 (3%)	
CHOP INTEND proportion with any worsening	5 (7%)	18 (49%)	
CMAP amplitude responders	26 (36%)	2 (5%)	$p=0.001$

CHOP INTEND - Children's Hospital of Philadelphia Infant Test of Neuromuscular Disorders; CI - confidence interval; CMAP - compound muscle action potential; HINE-2 - Module 2 of the Hammersmith Infant Neurological Examination
Note: HINE-2 responders were infants with a ≥2-point increase [or maximal score] in the ability to kick, OR ≥1-point increase in the motor milestones of head control, rolling, sitting, crawling, standing or walking, and improvement in more categories of motor milestones than worsening.

- Assumption that after adjustment for age, long-term mortality is the same in Gregoretti *et al*[31] and ENDEAR[14] is not plausible
- Uncertainty due to reconstruction of IPD from published Kaplan-Meier curve

(iii) Use of external data from Zerres *et al*[33] to inform later onset model

- Assumption that long-term mortality is the same as in CHERISH is not justified

(iv) Use of general population mortality

- Assumption that long-term mortality is systematically different between the studies and the general population (by assuming a constant HR) is not plausible

(v) Assumptions regarding treatment effect

- Description that a conservative HR of 1.0 is applied is misleading due to the implementation of the Type II adjustment

(vi) Concerns regarding SMA Type II adjustment

- No observed data to justify the use of Zerres *et al*[33] data or the adjustment factors used.

(i) Complexity of modelling approach

Jackson *et al*[49] present a framework for survival extrapolation using external data which is referenced by the company in justifying their approach (see clarification response,[2] question B9). If the external population has the same mortality at all times (or in the long-term) as that of the external population, then survival estimates from the external population can be used directly without adjustment. This assumption permits the direct use of data from Gregoretti *et al*[31] and Zerres *et al*[33] in the early onset and late onset models, respectively. Alternatively, OS may be assumed to be different, but systematically similar in such a way that the external data can be adjusted to estimate OS in the target population. This assumption permits the application of the adjusted general population mortality data. The validity of these assumptions is paramount to the reliability of the survival predictions; however, no clear justification for either assumption was presented by the company. The ERG considers that the plausibility of these assumptions is questionable and considers each case in further detail below.

Given the concerns regarding the use of external data, the ERG considers that a simpler approach based on extrapolating parametric models fitted to observed trial data may have been both more informative and more transparent than the approach adopted by the company. Consideration of appropriate external data is important; however, it could be used more simply to judge the plausibility of models fitted to observed data, or to inform certain parameters.[56] In their response to clarification questions from the ERG[2] (question B9), the company states that some parametric models provided plausible extrapolations (although they did not provide the best fit to the observed data) and so the ERG considers that using these may be a reasonable approach. Details of which models provided plausible predictions were not provided by the company.