Early Human Development 104 (2017) 17–21

Contents lists available at ScienceDirect

Early Human Development

journal homepage: www.elsevier.com/locate/earlhumdev

# Inter-observer agreement of the General Movements Assessment with infants following surgery

Cathryn Crowle [a,b,*], Claire Galea [c], Catherine Morgan [b,c], Iona Novak [b,c], Karen Walker [a,b,c], Nadia Badawi [a,b,c]

[a] Grace Centre for Newborn Care, The Children's Hospital Westmead, Sydney, Australia
[b] University of Sydney, Sydney, Australia
[c] Cerebral Palsy Alliance Research Institute, Sydney, Australia

ARTICLE INFO

*Article history:*
Received 26 April 2016
Received in revised form 2 November 2016
Accepted 13 November 2016

*Keywords:*
Reliability
Inter-observer agreement
General Movements Assessment
Prechtl
Infant

ABSTRACT

*Background:* The General Movements Assessment (GMA) is a validated and reliable method of identifying infants at risk of adverse neurodevelopmental outcomes, however there is minimal data available on the use of the GMA with infants following surgery.
*Aims:* The aim of this study was to investigate the inter-observer agreement for the GMA with this infant population.
*Study design:* Reliability and agreement study.
*Subjects:* This was a prospective cohort study of 190 infants (male n = 112) born at term (mean 38 weeks, SD 2 weeks).
*Outcome measures:* A GMA was conducted in the Neonatal Intensive Care Unit (NICU) following either cardiac surgery (n = 92), non-cardiac surgery (n = 93) or both types of surgery (n = 5), and then again at three months of age. All videos were independently assessed by three advanced trained clinicians. Agreement and reliability statistics were calculated between each pair.
*Results:* We found moderate to substantial levels of agreement in the writhing period (66–77%, $AC_1 = 0.53–0.69$). For fidgety general movements, agreement was classified as almost perfect, ranging from 86 to 89% ($AC_1 = 0.84–0.88$).
*Conclusions:* The GMA has high levels of inter-observer agreement when used with infants who have undergone surgery in the neonatal period, making it a valid, complementary assessment tool. Research is now underway to determine the ability of the GMA to predict neurodevelopmental outcomes in this population.

## 1. Introduction

Prechtl's Method on the Qualitative Assessment of General Movements (GMA) is a valid, reliable, non-invasive assessment tool for identifying infants at risk of poor neurodevelopmental outcomes, specifically cerebral palsy, with the aim of intervening early and improving outcomes. Whilst extensive evidence exists for use of the GMA, particularly with the preterm infant population [1–5] there is a paucity of published data on the use of the GMA with the infant surgical population, despite the known risk of later neurodevelopmental problems following surgery in the neonatal period [6].

We know from population-based data of infants in New South Wales that at one year of age, infants who have undergone surgery in the neonatal period demonstrated a delay (ranging from mild to severe) on all subscales of the Bayley Scales of Infant & Toddler Development [6]. For infants who had undergone cardiac surgery, the greatest difference was on the gross motor subscale, with 50% of the infants demonstrating a delay, compared with 20% of controls. Infants who underwent non-cardiac surgery performed worse in four of the five subtests (cognition, receptive language, fine motor and gross motor) with statistically significant lower mean scores than the control infants. Gross motor delay was evident in 38% of these infants, compared with 20% of the control group [6].

The GMA is an assessment of quality of movement and can be used with infants from birth until approximately 20 weeks post-term age. It involves video recording an infant's spontaneously generated movements as they lay supine in a quietly awake state, in order to evaluate the integrity of the infant nervous system. It is a clinically feasible tool to use with fragile post-surgical infants as it does not require any handling [7]. Movement quality is categorised by trained observers as outlined by Prechtl and colleagues [8]. From preterm age, up until approximately nine weeks post-term age, 'writhing' general movements are categorised as either normal, poor repertoire, cramped-

* Corresponding author at: Grace Centre for Newborn Care, The Children's Hospital at Westmead, Locked Bag 4001, Westmead 2145, Australia.
*E-mail addresses:* Cathryn.crowle@health.nsw.gov.au (C. Crowle), cgalea@cerebralpalsy.org.au (C. Galea), cmorgan@cerebralpalsy.org.au (C. Morgan), inovak@cerebralpalsy.org.au (I. Novak), Karen.walker@health.nsw.gov.au (K. Walker), nadia.badawi@health.nsw.gov.au (N. Badawi).

http://dx.doi.org/10.1016/j.earlhumdev.2016.11.001
0378-3782/© 2016 Elsevier Ireland Ltd. All rights reserved.

synchronised or chaotic. From nine to twenty weeks post-term age, 'fidgety' general movements are categorised as either normal, abnormal or absent. Normal general movements are associated with a normal outcome, whilst infants demonstrating persistently cramped-synchronised movements and/or absent fidgety movements, are at high risk of cerebral palsy [8].

The GMA is increasingly being used in clinical settings to identify infants who would benefit from targeted early intervention. We know the GMA has high reliability in the preterm infant population and those with Hypoxic Ischaemic Encephalopathy (HIE) [4,9]. Previous studies have reported inter-observer agreement rates ranging from 87 to 93% [10–12], and studies reporting on inter-observer reliability using a kappa statistic report fair to almost perfect agreement, with kappa between 0.36 and 0.94 [10, 12–17]. To date this remains unreported in the infant surgical population. As the profile of abnormal neurodevelopmental outcomes has different prevalence's in this surgical group, it is important to study this group separately, rather than assume rates of agreement will be the same. Just as high prevalence affects the predictive accuracy of cerebral palsy using the GMA [18], studies of inter-observer agreement are also affected by prevalence. Establishing strong inter-observer agreement between trained clinicians using the GMA with the infant surgical population is one important step in determining validity.

The aim of this study was to determine the inter-observer agreement and reliability of the GMA with infants who have undergone surgery in the neonatal period.

## 2. Methods

Procedures and reporting for this study followed the guidelines outlined by Kottner and colleagues in 2011; GRRAS (Guidelines for Reporting Reliability and Agreement Studies) [19]. This paper proposed the items which should be addressed when reliability and agreement studies are reported, such as the subject and rater population, the process for rating, the statistical analysis and the reporting of estimates of reliability and agreement.

### *2.1. Study design*

This was a prospective cohort study of 190 infants whose GMA videos were independently assessed by three advanced trained observers, two of whom were blinded to the infant's medical history. Levels of inter-observer agreement and reliability were measured.

### *2.2. Participants*

We conducted a prospective study on the use of the GMA with infants following surgery in the neonatal period, which recruited 304 infants from a level 6 NICU in NSW, Australia. Infants were enrolled if they required surgery within the first 30 days of life, and were eligible for follow-up in the development clinic. The development clinic sees infants >30 weeks gestation with congenital cardiac conditions, major surgical anomalies, or significant neurological problems. Written consent for GMs video assessment was gained from parents/carers and ethics approval for the study was obtained through The Sydney Children's Hospital Network, Human Research Ethics Committee. No family declined to participate.

From the study sample of 304 infants, 190 infants were eligible and included in analysis for this sub-study. We excluded the 82 infants who had missing data for the GMA in the writhing period, and nine infants where the GMA was completed outside the optimum age range to more reliably assess writhing or fidgety movements (refer to procedure below). We then omitted infants whose videos were scored as 'un-assessable' due to behavioural state (crying, fussing) ($n = 18$); infants who did not proceed to surgery ($n = 3$), and infants whose videos were not available to be viewed by all three observers ($n = 2$). This left a sample of 190 infants with writhing and fidgety videos that were able to be scored (illustrated in Fig. 1).

### *2.3. Observers*

The three observers had completed both the basic and advanced training courses offered by the GMs Trust. There were two occupational therapists and one physiotherapist, all experienced in the clinical use of the GMA. Two external observers, three years post-Advanced training certification, were blinded to the infant's medical history and any risk factors, and only provided with the age of the infant at the time of assessment, as recommended during GMA training. The third observer was the infant's treating clinician and could not be blinded to clinical details.

### *2.4. Procedure*

In order to ensure consistency with the procedure, a protocol was developed. This included guidelines for inclusion/exclusion of videos. Writhing videos needed to be longer than 2 min duration and taken prior to 46 weeks gestational age. From this age onwards, writhing movements involving complex, variable rotations along the axis of the limbs and through the trunk, slowly begin to disappear, as they are replaced by smaller, circular movements that are characteristic of the fidgety period [8]. By about nine weeks of age, fidgety movements become more apparent, and are at their peak by 12 weeks of age. Similarly, these movements fade out as voluntary, goal directed movements take over. Based on this, for the purposes of evaluating inter-observer agreement, fidgety videos needed to be taken between 10 and 15 weeks of age.

Videos were taken following the guidelines outlined by Prechtl and colleagues [8], which included infants lying on their back with minimal clothing, no dummy (pacifier) or toys, and in an adequate behavioural state to lie quietly whilst unwrapped. Videos were scored as 'un-assessable' (and excluded) when there was persistent crying or unsettled behaviour, despite several attempts.

Observers assessed the videos in independent locations, but were aware that their scores would be compared. There was a Gestalt setting of typical cases [8] prior to each rating period, which lasted a maximum time of 1 h. Rating categories were those described by Prechtl and colleagues [8]. In the writhing stage the categories were normal, poor repertoire (PR), cramped synchronised (CS) or chaotic; in the fidgety stage, categories were normal, abnormal or absent. For the analysis of writhing data, there were only five infants that were rated as having 'chaotic' general movements. A decision was made to omit these from the final analysis due to the very small number, and the very low prevalence across all populations [8]. This left 185 infants for the analysis of writhing GMA.

### *2.5. Statistical analysis*

Studies reporting on inter-observer agreement and reliability traditionally report the Cohen's kappa statistic, however when the prevalence of a condition is low, there is ample information available regarding the problems with this statistic [20,21]. For this study, an alternative reliability coefficient, the AC1 statistic was used, as it adjusts for chance agreement more appropriately than Cohen's kappa in this population [22,23]. However, Cohen's kappa (kappa) and percentage agreement were also calculated for inter-observer reliability. Linear weighted analysis for kappa was performed. Descriptive statistics were used to profile demographic characteristics of the sample.

Gwet's AC1 statistic and weighted kappa inter-observer agreement coefficients were interpreted using benchmark scales of Landis & Koch, and Altman [24] (refer to Table 1). 95% confidence intervals ($p = 0.05$) were calculated for weighted kappa and Gwet's AC1 statistic. Analysis was performed using Agree-Stat 2015 (Advanced Analytics,

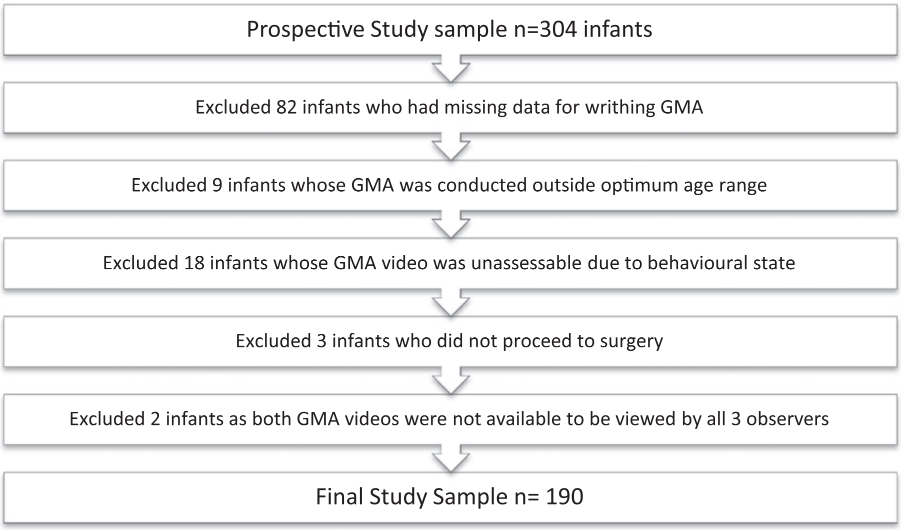
Prospective Study sample n=304 infants
Excluded 82 infants who had missing data for writhing GMA
Excluded 9 infants whose GMA was conducted outside optimum age range
Excluded 18 infants whose GMA video was unassessable due to behavioural state
Excluded 3 infants who did not proceed to surgery
Excluded 2 infants as both GMA videos were not available to be viewed by all 3 observers
Final Study Sample n= 190

**Fig. 1.** Method to obtain study sample.

LLC, Gaithersburg, Maryland) and Stata (StataCorp, 2015. Stata Statistical Software: Release 14. College Station, TX: StataCorp LP).

## 3. Results

In total, the GMAs for 190 infants were assessed by three independent observers in separate locations. The infants were mostly born at term (81%; mean = 38.1 weeks, SD 2.1 weeks) and had undergone either cardiac surgery (n = 92), non-cardiac surgery (n = 93) or both types of surgery (n = 5) in the first 30 days of life. Writhing GMAs were conducted whilst infants were in the NICU, at a mean gestational age of 40.3 weeks and predominantly post-surgery (mean = 9.65 days post-surgery, SD = 13.16). Fidgety GMAs were included as part of the first clinic follow-up appointment at three months of age (mean age = 12.2 weeks). Further infant characteristics are outlined in Table 2.

There was unanimous agreement between clinicians in 101 (54%) writhing GMAs and 155 (82%) of fidgety GMAs, with a majority agreement reached in all (100%) cases, i.e., at least two observers agreed. Agreement between pairs ranged from 66 to 76% in the writhing, and 86–89% in the fidgety (refer to Table 3). Table 4 outlines the actual GMs categories given by each pair of observers.

In the writhing period, moderate to substantial agreement was found between each pair of observers with comments noted by observers during rating periods indicative of even higher agreement. For example, for disagreement between normal and poor repertoire, the comments section often showed a comment of PR-normal. In the writhing period we found similar inter-observer agreement between the cardiac and non-cardiac surgical groups. The median $AC_1$ statistics were 0.60 (range 0.58 to 0.74) and 0.59 (range 0.50 to 0.63) respectively.

Almost perfect agreement was reported between each pair of observers in the fidgety period, with no significant difference between the surgical groups. For infants in the cardiac surgical group, the median $AC_1$ statistic was 0.86 (range 0.85 to 0.88) and the non-cardiac surgical group was 0.89, (range 0.85 to 0.90). The high prevalence of “normal” rating in the fidgety category clearly supported the use of the paradox-resistant $AC_1$ statistic. This is highlighted in the cardiac surgical group, where the inter-observer agreement between observer one and observer two with linear weighted kappa was fair (0.37) but with $AC_1$ it was almost perfect (0.86) which was reflective of the percentage agreement (0.87).

## 4. Discussion

In this study of inter-observer agreement of the GMA with infants following surgery, we obtained the following results: (a) inter-observer agreement between pairs as a percentage was high, with slightly better agreement in the fidgety period (b) inter-observer reliability calculated using the $AC_1$ statistic was also high, with moderate to substantial agreement in the writhing period and almost perfect agreement in the fidgety period (c) a majority agreement was reached between at least two observers in all assessments, which has clear clinical significance as a consensus rating could be reached.

The rates for inter-observer agreement reported in this study, based on percentage agreement and the $AC_1$ statistic, were similarly positive to those previously reported with other high risk infant populations [10,12–17]. Previous studies reported on the inter-observer reliability using a kappa statistic, but as mentioned there is ample information available regarding the problems with using the kappa statistic [20,21] as it can produce unexpected results under certain conditions known as the paradoxes of kappa [25]. The first paradox for kappa is that if the prevalence is high then the correction process will convert a high percentage agreement into a low kappa [21]. The high proportion of normal general movements in this population made it difficult to rely on Cohen's kappa as the best indicator of agreement. For example, agreement between observer one and two in the fidgety period was 86%, yet kappa was calculated to be 0.32, due to the high number of

**Table 1**
Benchmark scales for kappa and $AC_1$ statistic.

| | Altman | Landis & Koch |
|---|---|---|
| 0.00 to 0.20 | Poor agreement | Slight |
| 0.21 to 0.40 | Fair agreement | Fair |
| 0.41 to 0.60 | Moderate agreement | Moderate |
| 0.61 to 0.80 | Good agreement | Substantial |
| 0.81 to 1.00 | Very good agreement | Almost perfect |

**Table 2**
Characteristics of sample.

| | |
|---|---|
| Gestational age (mean, SD) weeks | 38.1, 2.1 |
| Birth weight (mean, SD) grams | 3048 (586) |
| Age at writhing assessment (mean, SD) weeks | 40, 2.3 |
| Age at fidgety assessment (mean, SD) weeks | 12, 1.4 |
| Male | 112 (59%) |
| Cardiac surgery | 92 (48%) |
| Non-cardiac surgery | 93 (49%) |
| Both types surgery | 5 (3%) |

**Table 3**
Rates of agreement between pairs of observers.

| | $AC_1$ statistic (CI) | Agreement | Strength Altman/Landis et al. | Weighted kappa (CI) |
|---|---|---|---|---|
| *Writhing* | | | | |
| Observer 1 & 2 | 0.53 (0.44–0.63) | 66% | Moderate/moderate | 0.37 (0.25–0.48) |
| Observer 1 & 3 | 0.69 (0.60–0.77) | 77% | Good/substantial | 0.56 (0.45–0.68) |
| Observer 2 & 3 | 0.58 (0.49–0.68) | 70% | Moderate/moderate | 0.46 (0.33–0.58) |
| *Fidgety* | | | | |
| Observer 1 & 2 | 0.84 (0.78–0.90) | 86% | Very good/almost perfect | 0.32 (0.12–0.52) |
| Observer 1 & 3 | 0.88 (0.83–0.93) | 89% | Very good/almost perfect | 0.50 (0.32–0.69) |
| Observer 2 & 3 | 0.88 (0.82–0.93) | 89% | Very good/almost perfect | 0.52 (0.34–0.71) |

infants rated as 'normal' (88%). The AC1 statistic has been shown to have better statistical properties when the extent of agreement is high in a situation of high trait prevalence [25], as in this study population. The AC1 statistic rendered a kappa value of 0.8752, because it adjusted for chance agreement more appropriately than Cohen's kappa [23].

Therefore it is not always appropriate to compare kappa between different studies or populations [20] because if one category is observed more in one study and not another then kappa will indicate a difference in inter-observer agreement which is not due to the observers but rather the study population. The second paradox is the lack of predictability of changes in kappa given changing marginal values [26]. Gwet's $AC_1$ statistic is an alternative to the unstable kappa and is paradox-resistant [24]. Gwet's $AC_1$ statistic adjusts for chance agreement more appropriately than kappa [23] and gives a less divergent perspective of inter-rater agreement [27].

Rates of agreement were slightly less between observer 2 and other clinicians. As levels of training and overall experience were similar, we propose that it is also important to consider frequency of use of the GMA in clinical practice. Observer 2 was using the GMA less frequently on a routine clinical basis prior to the commencement of the study.

Possible explanations for the slightly lower agreement in the writhing period could be the larger number of categories, and the age of the infants at the time of assessment in the writhing period, which in this study, was at term age. In a study of 700 observers of GMAs, it was reported that the term age infants were more difficult to assess than the preterm, or those three to five months post-term age [9].

Another factor to consider is the complex nature of the cases in this study, particularly in the writhing stage. Videos were not always typical examples of the writhing categories, or as clear as those used in training, an observation also reported by Bernhardt and colleagues [17], who reported fair to moderate rates of agreement in their study. In the present study it was difficult at times to distinguish between PR, PR-normal, and normal. In addition, whilst every attempt was made to ensure limbs were free of arm boards, lines and drains, at times this was not possible, perhaps interrupting the overall Gestalt perception. Furthermore, although attempts were made to blind the two external observers to the type of surgery, at times this was not possible due to the presence of a stoma bag, or operative scars.

Another possible limitation for this study is that infants were recruited from a single centre as a convenience sample. However, as this centre is a level 6 NICU that accepts the majority of state-wide referrals for major surgery, and almost all neonatal cardiac surgery, the study group should be representative of the wider infant surgical population.

To our knowledge, this is the first report on the inter-observer agreement of the GMA with infants post-surgery. In all assessments a majority percentage agreement between independent observers was achieved and the results confirm the high inter-observer agreement when used with infants following neonatal surgery. Team evaluation when interpreting the GMA is beneficial for improved accuracy, particularly for more difficult, non-standard cases. Ongoing regular clinical use of this tool is recommended, as well as professional development in the form of networks, such as the NSW network of GMs raters [28].

As intended, the GMA should be used in conjunction with other neurological and developmental assessments, imaging, and consideration of risk factors (multiple births, gestation, birth weight), in order to most effectively identify infants at risk of poor neurodevelopmental outcomes. This will allow early referral to specialised intervention services with the aim of improving overall developmental outcomes.

**Table 4**
GMs ratings between each pair of observers.

| Writhing | | | | | |
|---|---|---|---|---|---|
| Observer 1 | | Observer 2 | | | |
| | | Normal | Poor repertoire | CS | Total |
| | Normal | **47** | 15 | 3 | 65 |
| | Poor repertoire | 28 | **74** | 9 | 111 |
| | CS | 3 | 5 | **1** | 9 |
| | Total | 78 | 94 | 13 | 185 |
| Observer 1 | | Observer 3 | | | |
| | | Normal | Poor repertoire | CS | Total |
| | Normal | **50** | 14 | 1 | 65 |
| | Poor repertoire | 16 | **89** | 6 | 111 |
| | CS | 1 | 5 | **3** | 9 |
| | Total | 67 | 108 | 10 | 185 |
| Observer 2 | | Observer 3 | | | |
| | | Normal | Poor repertoire | CS | Total |
| | Normal | **48** | 28 | 2 | 78 |
| | Poor repertoire | 17 | **75** | 2 | 94 |
| | CS | 2 | 5 | **6** | 13 |
| | Total | 67 | 108 | 10 | 185 |
| Fidgety | | | | | |
| Observer 1 | | Observer 2 | | | |
| | | Normal | Abnormal | Absent | Total |
| | Normal | **155** | 3 | 11 | 169 |
| | Abnormal | 1 | **1** | 0 | 2 |
| | Absent | 11 | 1 | **7** | 19 |
| | Total | 167 | 5 | 18 | 190 |
| Observer 1 | | Observer 3 | | | |
| | | Normal | Abnormal | Absent | Total |
| | Normal | **157** | 0 | 12 | 169 |
| | Abnormal | 0 | **2** | 0 | 2 |
| | Absent | 7 | 1 | **11** | 19 |
| | Total | 164 | 3 | 23 | 190 |
| Observer 2 | | Observer 3 | | | |
| | | Normal | Abnormal | Absent | Total |
| | Normal | **155** | 1 | 11 | 167 |
| | Abnormal | 3 | **2** | 0 | 5 |
| | Absent | 6 | 0 | **12** | 18 |
| | Total | 164 | 3 | 23 | 190 |

Numbers in bold indicate the same rating given by both observers.

## References

[1] M. Burger, Q.A. Louw, The predictive validity of general movements-a systematic review, Eur. J. Paediatr. Neurol. 13 (5) (2009) 408–420.

[2] A.J. Spittle, L.W. Doyle, R.N. Boyd, A systematic review of the clinimetric properties of neuromotor assessments for preterm infants during the first year of life, Dev. Med. Child Neurol. 50 (4) (2008) 254–266.

[3] V. Darsaklis, L.M. Snider, A. Majnemer, B. Mazer, Predictive validity of Prechtl's method on the qualitative assessment of general movements: a systematic review of the evidence, Dev. Med. Child Neurol. 53 (10) (2011) 896–906.

[4] M. Bosanquet, L. Copeland, R. Ware, R. Boyd, A systematic review of tests to predict cerebral palsy in young children, Dev. Med. Child Neurol. 55 (5) (2013) 418–426.

[5] C. Einspieler, A.F. Bos, M.E. Libertus, P.B. Marschik, The general movement assessment helps us to identify preterm infants at risk for cognitive dysfunction, Front. Psychol. 7 (2016) 406.

[6] K. Walker, N. Badawi, R. Halliday, J. Stewart, G.F. Sholler, D.S. Winlaw, et al., Early developmental outcomes following major noncardiac and cardiac surgery in term infants: a population-based study, J. Pediatr. 161 (4) (2012) 748–752, e1.

[7] C. Einspieler, H. Yang, K.D. Bartl-Pokorny, X. Chi, F.-F. Zang, P.B. Marschik, et al., Are sporadic fidgety movements as clinically relevant as is their absence? Early Hum. Dev. 91 (4) (2015) 247–252.

[8] C. Einspieler, H.F.R. Prechtl, A.F. Bos, F. Ferrari, G. Cioni, Prechtl's Method on the Qualitative Assessment of General Movements in Preterm, Term and Young Infants, Mac Keith Press, London, 2004.

[9] T. Valentin, K. Uhl, C. Einspieler, The effectiveness of training in Prechtl's method on the qualitative assessment of general movements, Early Hum. Dev. 81 (7) (2005) 623–627.

[10] L. Adde, M. Rygg, K. Lossius, G.K. Øberg, R. Støen, General movement assessment: predicting cerebral palsy in clinical practise, Early Hum. Dev. 83 (1) (2007) 13–18.

[11] F. Ferrari, G. Cioni, C. Einspieler, M.F. Roversi, A.F. Bos, P.B. Paolicelli, et al., Cramped synchronized general movements in preterm infants as an early marker for cerebral palsy, Arch. Pediatr. Adolesc. Med. 156 (5) (2002) 460–467.

[12] Y. Noble, R. Boyd, Neonatal assessments for the preterm infant up to 4 months corrected age: a systematic review, Dev. Med. Child Neurol. 54 (2) (2012) 129–139.

[13] D.M. Romeo, A. Guzzetta, M. Scoto, M. Cioni, P. Patusi, D. Mazzone, et al., Early neurologic assessment in preterm-infants: integration of traditional neurologic examination and observation of general movements, Eur. J. Paediatr. Neurol. 12 (3) (2008) 183–189.

[14] V. van Kranen-Mastenbroek, R. van Oostenbrugge, L. Palmans, A. Stevens, H. Kingma, C. Blanco, et al., Inter- and intra-observer agreement in the assessment of the quality of spontaneous movements in the newborn, Brain Dev. 14 (5) (1992) 289–293.

[15] A.F. Bos, A.J. van Loon, M. Hadders-Algra, A. Martijn, A. Okken, H.F.R. Prechtl, Spontaneous motility in preterm, small-for gestational age infants II. Qualitative aspects, Early Hum. Dev. 50 (1) (1997) 131–147.

[16] G. Cioni, A.F. Bos, C. Einspieler, F. Ferrari, A. Martijn, P.B. Paolicelli, et al., Early neurological signs in preterm infants with unilateral intraparenchymal echodensity, Neuropediatrics 31 (5) (2000) 240–251.

[17] I. Bernhardt, M. Marbacher, R. Hilfiker, L. Radlinger, Inter- and intra-observer agreement of Prechtl's method on the qualitative assessment of general movements in preterm, term and young infants, Early Hum. Dev. 87 (9) (2011) 633–639.

[18] S.E. Hanna, High and variable prevalence does matter in reviews of diagnostic accuracy in cerebral palsy, Dev. Med. Child Neurol. 55 (5) (2013) 397–398.

[19] J. Kottner, L. Audige, S. Brorson, A. Donner, B.J. Gajewski, A. Hróbjartsson, et al., Guidelines for reporting reliability and agreement studies (GRRAS) were proposed, Int. J. Nurs. Stud. 48 (6) (2011) 661–671.

[20] A. Viera, J. Garrett, Understanding interobserver agreement: the kappa statistic, Fam. Med. 37 (5) (2005) 360–363.

[21] A.R. Feinstein, D.V. Cicchetti, High agreement but low kappa: I. The problems of two paradoxes, J. Clin. Epidemiol. 43 (6) (1990) 543–549.

[22] N. Wongpakaran, T. Wongpakaran, D. Wedding, K.L. Gwet, A comparison of Cohen's kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples, BMC Med. Res. Methodol. 13 (1) (2013) 1.

[23] H.H. Nishiura, A robust statistic $AC_1$ for assessing inter-observer agreement in reliability studies, Nippon Hōshasen Gijutsu Gakkai Zasshi 66 (11) (2010) 1485–1491.

[24] K.L. Gwet, Handbook of Inter-rater Reliability, Advanced Analytics Press, Maryland, USA, 2014.

[25] K.L. Gwet, Computing inter-rater reliability and its variance in the presence of high agreement, Br. J. Math. Stat. Psychol. 61 (1) (2008) 29–48.

[26] C.A. Lantz, E. Nebenzahl, Behavior and interpretation of the κ statistic: resolution of the two paradoxes, J. Clin. Epidemiol. 49 (4) (1996) 431–434.

[27] E.P. Mulligan, D.Q. McGuffie, K. Coyner, M. Khazzam, The reliability and diagnostic accuracy of assessing the translation endpoint during the Lachman test, Int. J. Sports Phys. Ther. 10 (1) (2015) 52–61.

[28] C. Morgan, C. Crowle, T.A. Goyen, C. Hardman, M. Jackman, I. Novak, et al., Sensitivity and specificity of General Movements Assessment for diagnostic accuracy of detecting cerebral palsy early in an Australian context, J. Paediatr. Child Health 52 (1) (2016) 54–59.