Full Length Article

JHS(E)

Journal of Hand Surgery
(European Volume)
2022, Vol. 47(10) 1016–1024

DOI: 10.1177/17531934221107264
journals.sagepub.com/home/jhs


Inter- and intra-rater reliability of the Oberg–Manske–Tonkin classification of congenital upper limb anomalies

Ida Neergård Sletten[1], Mona Irene Winge[1], Wiebke Hülsemann[2], Marianne Arner[3,4], Karina Liv Hansen[5] and Jarkko Jokihaara[6,7]

Abstract

On two occasions, five surgeons classified a cohort of 150 consecutive patients with congenital upper limb anomalies according to the Oberg–Manske–Tonkin classification (2020 update). We estimated reliability for the main anomaly code by means of Cohen's kappa (K) for ten rater pairs for five common and easily distinguishable anomalies (Group 1), and for all the other anomalies (Group 2). Inter-rater reliability for all patients ($n = 150$) was substantial, almost perfect for Group 1 ($n = 64$), but only moderate for Group 2 ($n = 86$). Intra-rater reliability was higher for all groups. We suggest simplifications to the Oberg–Manske–Tonkin classification and highlight specific requirements for instructions to increase its reliability.

Level of evidence: I

Keywords

Classification, congenital anomaly, Oberg–Manske–Tonkin, OMT, reliability

Date received: 14th March 2022; Revised 27th May 2022; accepted: 27th May 2022

Introduction

After two separate publications suggesting a better classification for congenital upper limb anomalies (CULAs) (Manske and Oberg, 2009; Tonkin, 2006), Oberg, Manske and Tonkin together published their OMT classification (Oberg et al., 2010). Tonkin and co-workers (2013) reported the utility of the OMT classification in an Australian patient cohort and suggested a refined version, which hand surgeons in Sweden used in an epidemiological study (Ekblom et al., 2014). The International Federation of the Societies of Surgery of the Hand (IFSSH) adopted a slightly modified version of this as their recommended CULA classification in 2014 (Ezaki et al., 2014). This was shortly thereafter used in an epidemiological study in the USA (Goldfarb et al., 2015), which led to a 2015 IFSSH update that included the American study's minor suggested additions (Tonkin and Oberg, 2015). The IFSSH has recommended a review of the OMT classification every third year (Ezaki et al., 2014), and published a new update in 2020 (Goldfarb et al., 2020). This update aimed to improve the classification terminology and

[1]Division of Orthopaedic Surgery, Oslo University Hospital, Oslo, Norway
[2]Department of Hand Surgery, Catholic Children's Hospital Wilhelmstift, Hamburg, Germany
[3]Department of Hand Surgery, Södersjukhuset, Stockholm, Sweden
[4]Department of Clinical Science and Education, Södersjukhuset, Karolinska Institutet, Stockholm, Sweden
[5]Department of Orthopedic Surgery, Odense University Hospital, Odense, Denmark
[6]Department of Hand Surgery, Tampere University Hospital, Central Hospital, Tampere, Finland
[7]Faculty of Medicine and Health Technology, Tampere University, Finland

Corresponding Author:
Ida Neergård Sletten, Division of Orthopedic Surgery, Oslo University Hospital, Postboks 4950, Nydalen, 0424 Oslo. Norway.
Email: ida.sletten@icloud.com

to make the classification options clearer (e.g. symbrachydactyly versus transverse deficiency). The authors added a few phenotypes, removed one and moved some within the system (e.g. cleft hand and congenital contractures). As yet, no one has published a reliability study of the 2020 version of the OMT. One study has shown substantial inter-rater reliability and almost perfect intra-rater reliability of the previous 2015 version (Bae et al., 2018), and only one reliability study originates from outside the institutions to which the developers of the OMT classification were affiliated (Uzun et al., 2020).

The CULA North project is an umbrella term for five newly established CULA registries in Denmark, Finland, Germany, Norway and Sweden. The registries have separate databases, but the registry developers have created a common prospective data collection protocol to provide comparable data for future research purposes. A reliable CULA classification is fundamental for such registry data. The CULA North Oslo Registry (Norway) was established in 2018 as the first of the five registries, and we have so far included more than 600 patients. The registry is based on the OMT classification, and we have noted that its coding can be ambiguous and is open for discussion. We do consider the 2020 OMT classification update to be an improvement (Hülsemann et al., 2020), but its reliability in systematic clinical use is unknown. Hence the aim of this study was to test the inter- and intra-rater reliability of the OMT 2020 on a cohort of consecutive registry patients from a single institution.

Methods

We have conducted this methodological study in accordance with the Quality Appraisal of Diagnostic Reliability (QUAREL) tool (Lucas et al., 2010). An independent statistician performed a power analysis using the standard error of kappa in a previous reliability study (Bae et al., 2018), and found that a sample size of between 145 and 205 patients was required to investigate agreement (between the raters), believed to be from 0.85 to 0.90 using a 95% confidence interval of width 5%. Based on the power calculation, the study sample was a consecutive cohort of the first 150 CULA patients included in the CULA North Oslo Registry between May 2018 and October 2019. Three more patients were included in the registry during the time interval (two patients with floating accessory ulnar fingers and one with small finger camptodactyly), but they were excluded from this study because we lacked radiographs and photos. Oslo University Hospital is Norway's largest hospital. We treat most of the CULA patients in our Regional Health Trust, which includes 60% of Norway's population. All CULA patients are included at their first visit, except those with OMT Type IIIB (dysplasias – tumorous conditions). For this study, the registry developer (INS) created individual patient presentations of 150 consecutive patients based on their medical history and details from clinical examinations. For all, radiographs of both upper limbs were included, and also, for most patients, photos of the affected limb(s). For the second round of classification 2 months later, INS randomized the order of patient presentations. INS instructed the raters not to discuss the classification or the patients during the study period, and not to review any details of the cohort between the two classification rounds.

Raters

The five raters (MIW, WH, MA, KLH and JJ) were specialists in orthopaedic, plastic or hand surgery, all leading the CULA treatment in their institutions in five different European Union/European Economic Area countries. They had 8 to 40 years of experience in hand surgery, and 5 to 30 years of CULA surgery experience. Four raters worked in university hospitals and one in a specialized children's hospital, all hospitals with $>$65 new CULA patients per year. Two of the raters were university professors in hand surgery, and two were members of the international Congenital Hand Anomaly Study Group (CHASG).

Coding

INS instructed the raters to read all the relevant publications on OMT classification use (Bae et al., 2018; Ekblom et al., 2014; Ezaki et al., 2014; Goldfarb et al., 2015, 2020; Manske and Oberg, 2009; Oberg et al., 2010; Tonkin, 2006; Tonkin and Oberg, 2015; Tonkin et al., 2013; Uzun et al., 2020) and to use the codes and text from the IFSSH homepage (www.ifssh.info/scientific_committee_reports.php), because the OMT 2020 update article contains some errors (Goldfarb et al., 2020). To mirror the use of the classification in clinical practice, we purposely did not arrange a consensus meeting or give any other instructions to the raters on how to use the OMT 2020 classification.

For each patient, the raters entered a ranked list of anomaly codes and standardized text according to OMT 2020 with specification of the right or left side, in order from the most to least important (Tonkin et al., 2013). They were instructed to use OMT 2020 group headings if the anomalies were impossible to sub-classify. Thus, in total 99 OMT anomaly codes were available in main Groups I–III (malformations,

deformations and dysplasias), and 51 codes in main Group IV (syndromes).

Statistical methods

We estimated inter-rater reliability for the primary anomaly codes by kappa statistics to correct for agreements that could be explained by chance (Hallgren, 2012; McHugh, 2012; Sim and Wright, 2005). As there were 99 possible anomaly codes, we expected agreement by chance to be low, and therefore also calculated percentage agreement among the raters. We used Cohen's kappa for rater pairs because our study design was fully crossed, and Fleiss' multi-rater kappa cannot be used unless there are non-unique, randomly selected raters (Hallgren, 2012). For the inter-rater analyses, all the five raters were paired with each other, yielding ten rater pairs in total. Percentage agreement and Cohen's kappa were calculated for each pair. From the outcome of all the ten rater pairs, mean values of percentage agreement and Cohen's kappa with 95% confidence intervals were calculated (Hallgren, 2012; Huhnstock et al., 2017). For intra-rater analyses, the same group means were calculated from the self-agreement for each rater at the two time points. Variants of Cohen's kappa may be selected based on problems of prevalence and bias in the marginal distributions (Hallgren, 2012; Sim and Wright, 2005), but it was considered too complex to estimate in this study. Possible kappa values range from –1 to 1, where –1 indicates no agreement, 0 random agreement and 1 perfect agreement. We interpreted values between 0 and 1 as suggested by Landis and Koch (1977a), with $<$0.20 as poor, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial and $>$0.80 as almost perfect agreement. We regarded kappa values lower than 0.60 as a clear indication of inadequate agreement (McHugh, 2012). We recorded classification reliability for all patients ($n = 150$), as well as separately for patients with a common and easily distinguishable anomaly (Madelung deformity, radial or ulnar polydactyly, cutaneous syndactyly or isolated camptodactyly; Group 1), and for all the other patients (Group 2).

We compared the number of anomaly codes per patient with the intraclass correlation coefficient (ICC) (Hallgren, 2012; Sim and Wright, 2005). The interpretation of ICC values followed the guidelines applied with kappa statistics (Landis and Koch, 1977a).

To identify problematic anomaly codes, we established *majority agreement* on the classification for each patient when at least three raters used the same primary (ranked as the most important) code. As agreement between three out of five raters is equivalent to agreement in only three out of ten rater pairs, we defined *strong agreement* when at least four raters agreed (equivalent to agreement in six out of ten rater pairs). If majority agreement was reached, we counted the number of patients with each OMT anomaly code, and recorded the type of the majority agreement (perfect: 5–0 if all five raters agree; split: 4–1, 3–2 or 3–1–1) (Landis and Koch, 1977b). If majority agreement on the primary code could not be obtained (raters split 2–2–1, 2–1–1–1 or 1–1–1–1–1), the patients were analysed separately. In an additional analysis, the other (secondary) codes were used to raise the level of agreement from majority to strong, but not to create majority agreement if fewer than three raters had used the same primary code.

Ethical aspects

The Oslo University Hospital's Data Protection Officer approved the study, and the Regional Health Trust's Ethical Committee considered the project as a quality study. All study patients or their caregivers gave a written consent at inclusion in the CULA North Oslo Registry for usage of their registry data in future research studies. It was not possible to identify any of the patients from the photos.

Results

Inter-rater reliability

Mean percentage agreements and mean kappa estimates with confidence intervals are presented in Table 1 and the data for each of the ten rater pairs are presented in Supplementary Table S1 (available online). Agreement was high on the number of codes per patient (Table 2). The raters varied in how often they used unspecified codes (group headings). If a rater had not been able to apply even a group heading code for a patient, the code 'not applicable' (NA) was used in the statistical analyses (Table 2). No patient was ever considered as unclassifiable (NA) by more than one rater at a time. There were no indications from the rater pair analyses that the most senior CULA surgeons agreed more with each other than with the younger surgeons (Supplementary Table S1, available online).

The raters had a strong majority agreement on the primary anomaly code in 74% of the patients in Round 1, and in 73% in Round 2 (Table 3). These numbers increased to 83% and 82% when we also included the secondary codes.

Table 1. Mean inter-rater reliability on main anomaly code for the ten rater pairs.

Readings	Round	Percentage of agreement, mean (95% CI[a])	Cohen's kappa, mean (95% CI[b])	Agreement[c]
All ($n = 150$)	1	71% (67 to 76%)	0.70 (0.66 to 0.74)	Substantial
	2	69% (64 to 73%)	0.67 (0.62 to 0.72)	Substantial
Group 1 ($n = 64$)	1	92% (90 to 95%)	0.90 (0.87 to 0.94)	Almost perfect
	2	84% (78 to 89%)	0.80 (0.73 to 0.86)	Almost perfect
Group 2 ($n = 86$)	1	56% (50 to 62%)	0.54 (0.48 to 0.60)	Moderate
	2	58% (52 to 63%)	0.56 (0.50 to 0.61)	Moderate

[a]95% confidence intervals for mean percentage agreement for rater pairs.
[b]95% confidence intervals for mean kappa values for rater pairs.
[c]According to Landis and Koch (1977a).
Group 1. Common and easily distinguishable anomalies: IA2vii Madelung deformity ($n = 11$), IB2iii radial polydactyly ($n = 22$), IB2vi ulnar polydactyly ($n = 10$), IB4ia cutaneous (simple) syndactyly ($n = 7$) and IIIC2i isolated camptodactyly ($n = 14$).
Group 2. All other anomalies.

Table 2. Inter- and intra-rater agreement on number of Oberg–Manske–Tonkin anomaly codes for each patient.

Round	Rater	*n*	Median (range)	Mean (95% CI)	ICC (95% CI)
1	1	150	2 (1–9)	2.2 (2.0 to 2.4)	0.90 (0.87 to 0.92)
	2	149	2 (0–6)	1.9 (1.7 to 2.1)	
	3	144	1 (0–4)	1.6 (1.4 to 1.7)	
	4	148	2 (0–7)	1.9 (1.7 to 2.1)	
	5	150	2 (1–5)	1.9 (1.8 to 2.1)	
2	1	150	2 (1–9)	2.2 (2.0 to 2.4)	0.90 (0.86 to 0.93)
	2	150	2 (1–7)	2.0 (1.8 to 2.1)	
	3	144	1 (0–4)	1.5 (1.3 to 1.6)	
	4	148	2 (0–7)	1.8 (1.7 to 2.0)	
	5	150	2 (1–5)	1.9 (1.8 to 2.1)	

n: number of patients where the rater gave at least one code; 95% CI: 95% confidence interval; ICC: intraclass correlation coefficient.

Malformations – Proximal-distal axis. The raters disagreed on whether anomalies were symbrachydactylies or transverse deficiencies, also in patients who had nubbins with nails. They also disagreed on the coding for patients with short forearms and symbrachydactyly or transverse deficiency distal to the wrist. Some used a Malformation IA or a IB code, and some used both codes in a non-consistent order. Additionally, they disagreed on whether anomalies should be classified as brachydactyly or clinodactyly, or as brachydactyly or symbrachydactyly.

Malformations – Radial-ulnar (anterior-posterior) axis. The raters usually agreed that a patient had radial longitudinal deficiency (RLD) but disagreed on whether to use IA or IB codes and in which order, if both were used. They disagreed similarly about patients with ulnar longitudinal deficiency (ULD) with shorter, otherwise normal forearms (ULD Type 0) (Havenhill et al., 2005).

Malformations – Unspecified axis. The raters agreed perfectly on patients with cutaneous syndactyly, but not regarding osseous syndactyly, syndromic syndactyly and synpolydactyly. The raters used the codes for the latter three variably, also for patients who other raters had classified as hand ULD or cleft hand. Furthermore, they did not agree whether abnormal shoulder muscles should be added as an additional code in Poland syndrome and Sprengel's deformity.

Deformations and dysplasias. The raters had a strong agreement on most patients with deformations or camptodactyly, but not on patients with other congenital contractures. The raters all chose a main code from subgroup IIIC (except camptodactyly) in five patients. For two out of these five patients, the raters had a majority agreement on the same anomaly code in both rounds. For three of the patients, the rates had a majority agreement in only one of the rounds.

No agreement. For 13 patients in Round 1 and for 18 in Round 2, fewer than three raters agreed on the primary code. Among these, for eight patients in

Table 3. Raters' majority agreement on main anomaly code given as number of patients per Oberg–Manske–Tonkin classification code.

		Type of raters' majority agreement							
		All agreed 5–0		Split 4–1		Split 3–2 or 3–1–1		≥3 agreed (total)	
Code	Text	R1	R2	R1	R2	R1	R2	R1	R2
IA1iia	Symbrachydactyly a) Poland			1	1		2	1	3
IA1iib	Symbrachydactyly b) excluding Poland			3	3	3	1	6	4
IA1iiib	Transverse deficiency b) segmental	1						1	
IA2i	Radial longitudinal deficiency	1	1		1	3		4	2
IA2ii	Ulnar longitudinal deficiency					1	1	1	1
IA2iv	Radiohumeral synostosis			1	1			1	1
IA2v	Radioulnar synostosis	3	2	3	4	1		7	6
IA2vi	Congenital dislocation of the radial head	2	2					2	2
IA2vii	Forearm hemi-physeal dysplasia	10	1		9	1	1	11	11
IB1i	Brachydactyly	4	9	6	2	1	1	11	12
IB1ii	Symbrachydactyly hand plate	2		1	3	1	2	4	5
IB1iii	Transverse deficiency hand plate			1	2	2		3	2
IB1iv	Cleft hand	1	1			1		2	1
IB2i	RLD, hypoplastic thumb		1	4	5	3	3	7	9
IB2ii	ULD, hypoplastic ulnar ray			1		3	4	4	4
IB2iii	Radial polydactyly	19	19	2	1	1	2	22	22
IB2iva	Triphalangeal thumb a) five finger hand	1	1					1	1
IB2vi	Ulnar polydactyly	8	8	2			1	10	9
IB4ia	Cutaneous (simple) syndactyly	6	6			1		7	6
IB4iia	Osseous (complex) syndactyly	1	1	1	1	1	1	3	3
IB4iib	Clinodactyly			3	3			3	3
IB4iiia	Syndromic syndactyly (e.g. Apert hand)		1	1				1	1
IB4iiib	Synpolydactyly	1	1			1		2	1
IIA	Constriction ring sequence	2	3	1				3	3
IIIA1i	Hemihypertrophy	1	1					1	1
IIIA2i	Macrodactyly	1	1			1	1	2	2
IIIC1i	AMC- Amyoplasia	1	1				1	1	2
IIIC1ii	AMC- Distal arthrogryposis			1	1	1		2	1
IIIC2i	Camptodactyly	11	12	3			1	14	13
IIIC2ii	Thumb in palm						1	0	1
	Total, *n*	76	72	35	37	26	23	137	132
	Total, proportion of all patients, %	51%	48%	23%	25%	17%	15%	91%	88%
	Strong agreement; ≥4 raters, *n* (%)	R1: *n* = 111 (74%) R2: *n* = 109 (73%)							
	<3 raters agreed; *n*							13	18
	Total							150	150

R1: round 1; R2: round 2; RLD: radial longitudinal deficiency; ULD: ulnar longitudinal deficiency; AMC: amyoplasia multiplex congenita. The shaded lines represent the five supposedly easily distinguishable and commonly occurring anomalies (Group 1).

Round 1 and for seven in Round 2, the anomaly seemed difficult to classify (e.g. IA symbrachydactyly versus IA transverse deficiency, IB symbrachydactyly versus cleft hand, hand ULD, complex syndactyly or synpolydactyly). For five patients in Round 1 and 11 in Round 2, the raters seemed to agree on the anomaly, but they used different codes (IA versus IB codes, different IIIC codes, unspecific heading versus specific code under it). For seven patients, majority agreement was not reached in either round (four patients with an apparent anomaly classification problem and three patients with a code problem).

Intra-rater agreement

Percentage agreement and kappa estimates are presented in Table 4 and in Supplementary Table S2 (available online). The mean intra-rater ICC for the number of anomaly codes for each patient was 0.88 (range 0.68–0.98).

Table 4. Mean intra-rater reliability on main anomaly code for the five raters.

Readings	Percentage of agreement, mean (95% CI[a])	Cohen's kappa, mean (95% CI[b])	Agreement[c]
All (n = 150)	84% (74 to 94%)	0.83 (0.72 to 0.94)	Almost perfect
Group 1 (n = 64)	92% (82 to 100%)	0.90 (0.77 to 1.00)	Almost perfect
Group 2 (n = 86)	77% (67 to 88%)	0.76 (0.65 to 0.87)	Substantial

[a]95% confidence intervals for mean percentage intra-agreement for the five raters.
[b]95% confidence intervals for mean kappa values for the five raters.
[c]According to the criteria of Landis and Koch (1977a).
Group 1. Common and easily distinguishable anomalies: IA2vii Madelung deformity ($n = 11$), IB2iii radial polydactyly ($n = 22$), IB2vi ulnar polydactyly ($n = 10$), IB4ia cutaneous (simple) syndactyly ($n = 7$) and IIIC2i, isolated camptodactyly ($n = 14$).
Group 2. All other anomalies.

Discussion

The overall inter-rater reliability of the OMT classification was substantial in our study, in agreement with previous studies (Bae et al., 2018; Uzun et al., 2020). However, this was apparently mainly due to the five easily distinguishable and common anomalies (Group 1) for which the mean inter- and intra-rater reliabilities were almost perfect. The inter-rater reliability for the larger Group 2 of all the other anomalies was only moderate with mean kappa values below 0.60. In general, our kappa values were only slightly lower than the unadjusted percentage agreement. This finding verified our assumption that agreement by chance was low, and thereby validated our unadjusted majority agreement calculations.

The main strength of our study is the high internal and external validity. To ensure high internal validity the raters had not communicated about details in the OMT before the study and had no contact during the study. Furthermore, we minimized recall bias by having a 2-month interval between the sessions and minimized code translation bias as the raters added standardized text to all codes. We chose five raters in order to establish a majority agreement for each patient.

To achieve high external validity, both the patients and the raters were representative of the practice for which the classification was intended. The cohort of CULA patients in our study can be considered representative of the patient population in the Nordic countries. By including them consecutively, we minimized inclusion selection bias that can otherwise give too favourable or unfavourable reliability outcomes, depending on the level of difficulty in classifying the anomalies. The raters had varying levels of expertise, as in a real-life setting. They only used published OMT classification material and did not use any unauthorized local guidelines. In contrast, Bae and co-workers (2018) held a 'consensus-building exercise' before the classifications, and Uzun and co-workers (2020) reported that their raters 'received education' without giving further information. If we had given unauthorized rater guidance or used consensus meetings, the inter-rater reliability would have probably been higher for Group 2 in our study, but doing so would have weakened the study's external validity. We made this important methodological decision early in the planning phase, as we wanted to test the reliability in a setting as close to every-day clinical practice as possible. In a clinical study, the study protocol may give specific instructions on how to use the OMT classification to identify the CULAs in question. If such instructions are not universally acknowledged, they might affect the generalizability of the study results. To ensure useful study outcomes for clinical practice and valid comparisons between clinical studies, a CULA classification used to define inclusion criteria should be the same as that used for diagnosing patients. Furthermore, the classification should be so unambiguous that local interpretations and guidelines are unnecessary.

The main limitations in our study were the number of patients included and the number of raters. Including more patients would have increased the number of rare anomalies, but the proportion of common and easily distinguishable anomalies would probably have been the same. Adding more raters might have provided more accurate reliability estimates, but we considered that the growing complexity of the study would have outweighed the benefit. Furthermore, assessment of medical history and photos cannot give as much information as a real-life clinical examination. This was the best solution achievable in our setting, as using raters from different countries was considered to be methodologically best.

The number of anomaly codes per patient was higher in our study than in a Dutch cohort where 21% of the patients received more than one OMT code (Baas et al., 2018). In our study, majority

agreement increased to strong agreement in 10% of the patients when we included secondary codes, indicating that multiple codes are beneficial. One can argue that the code order is irrelevant if all are entered, but only if the main aim is to identify all patients with a certain code. Nevertheless, for most applications a classification is assumed to categorize by the most relevant factor (primary code). However, code ranking can be difficult because choosing which code is the most important may depend on patient age, activities or other preferences, or may be based just on the physician's assumption.

We chose not to analyse codes for both limbs separately in bilateral CULAs even though this would have increased the number of CULAs. By doing so, the limbs would not have been statistically independent. Including several codes from the same patient would have raised multiplicity issues and would have weakened the analyses. Hence, the most important CULA (primary code) was included for each patient in the kappa analyses.

Similar to our findings, Bae and co-workers (2018) demonstrated that certain types of anomalies have higher inter-rater reliability than others. For many of our patients it seemed that the raters agreed on the phenotype, but not on its coding. For example, the division of the I Malformation into IA and IB is in our opinion somewhat artificial, at least for symbrachydactyly, transverse deficiency, RLD and ULD. The first OMT classification included these four phenotypes under IA (Oberg et al., 2010), but they were later split into IA and IB types (Tonkin et al., 2013). Nevertheless, it was emphasised that symbrachydactyly usually includes a proximal component of limb deficiency (Tonkin et al., 2013). In addition, most patients with thumb hypoplasia have carpal abnormalities and the radiological differentiation between RLD types N and 0 cannot be made before the child is 8 years of age (James et al., 1999). In the 2020 update, it is not clear whether the hand plate codes include the carpus. All patients with RLD Types 1–4 have thumb hypoplasia (James et al., 1999), thus adding an IB code to a patient with RLD Type IA can be considered to be unnecessary (Tonkin et al., 2013). These debatable instructions have led to individual practice guidelines. For example, the Operation Manual from the American Congenital Upper Limb Differences (CoULD) Registry (founded in 2014) instructs that both IA and IB codes should be applied in RLD (https://kidshandregistry.com). This approach seems beneficial because the manual has incorporated the Manske classification into the IB code (Manske et al., 1995). There are other inconsistencies in the IA versus IB distinction in the OMT, including that some IA malformations do not affect the entire limb (e.g. radioulnar synostosis), and some IB malformations commonly also have proximal involvement of their limbs (e.g. Apert hand).

Despite the clear instruction in the OMT 2020 on the difference between symbrachydactyly and transverse deficiency, many patients with transverse defects proximal to the wrist level with ectodermal elements were classified as having transverse deficiencies. This reflects the view that CULA surgeons may consider symbrachydactyly to be a hand anomaly. We have noted in our CULA North Oslo registry that ectodermal elements are present in most patients with transverse reduction defects, in accordance with the findings of Kallemeier and co-workers (2007). The distinction between symbrachydactyly and transverse deficiency is meaningful in the context of microsurgical toe-transfer for grip reconstruction. Especially in transverse reduction defects above the wrist, the difference between the two diagnoses might be too unclear to be a main divider in the OMT classification.

More surprisingly than the above-mentioned disagreements, we found that our raters coded IB1 and IB4 malformations differently. They classified some hands with osseous syndactyly as symbrachydactyly, osseous syndactyly, syndromic syndactyly or synpolydactyly, and this might suggest that not all CULAs fit into the classic descriptions. Furthermore, there are no instructions on whether syndromic syndactyly applies to Apert hand only, or to any patient with syndactyly and a syndrome (e.g. patients with Down's syndrome and syndactyly).

The raters did not agree on the use of clinodactyly versus brachydactyly, even though by definition, clinodactyly means lateral deviation and brachydactyly means shortness of a digit. The most common types of brachydactyly are A2 and A3 (Temtamy and McKusick, 1978), and include clinodactyly in most patients. It is unclear whether clinodactyly can appear without brachydactyly.

It is sensible to use as few codes as possible in CULA registries to avoid redundancy and overlapping. By definition, Leri-Weill dyschondrosteosis includes Madelung deformity, and Apert syndrome includes Apert hands. A previous OMT classification gave the instruction to use only Group IV syndrome codes for patients with a known syndrome (Tonkin et al., 2013), but it was updated to use specific anomaly codes with syndrome codes without giving advice on how to rank them (Tonkin and Oberg, 2015). In our study, most raters applied both, and ranked specific anomaly codes higher than syndrome codes.

The 2020 OMT classification has high reliability for easily distinguishable anomalies. For more complex anomalies, the outcome in our study indicates that

the classification might be too complicated. We believe that simplifications should increase the clinical applicability. Merging the arm (IA) and hand plate (IB) malformation subgroups and merging symbrachydactyly and transverse deficiencies should be considered. We also suggest removing, or better defining, IB4iib clinodactyly, changing IB4iiia syndromic syndactyly back to IB4iiia Apert hand, and removing the unspecific codes IB4iiic and IIB. We suggest that subgroups under IIIC congenital contractures could be simplified, for example, to IIIC1 amyoplasia multiplex congenita, IIIC2 distal arthrogryposis, IIIC3 isolated camptodactyly and IIIC4 isolated thumb-in-palm deformity. Also, we suggest that the use of Group IV syndrome codes should be better described.

A more unambiguous and 'user-friendly' CULA classification could be very useful both in clinical work and in research on CULA patients. We propose that the IFSSH simplifies the classification and presents detailed user instructions with it in the next update, to increase its inter-rater reliability. Comprehensive work aiming to assist a practical user is ongoing through the creation of OMT mobile apps (Lam, 2019a, 2019b). We strongly support the further refinement of this important project, preferably based on additional reliability studies or consensus reports. For the time being, the CULA North has, with permission from the American CoULD Registry, decided to follow their Operation Manual for classification.

Acknowledgements Thanks to statistician Lien My Diep at Oslo Centre for Biostatistics and Epidemiology, Oslo University Hospital, Oslo, Norway for performing the power analysis and for statistical advice. Thanks to Norwegian Professional Network for Congenital Limb Anomalies for supporting the writing of this article with a research grant.

Declaration of conflicting interests The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: INS received a personal research grant from Norwegian Professional Network for Congenital Limb Anomalies (no grant number available) to cover a 2-month research leave, of which two weeks were spent to finish this article. The Network had no role in the design of the study, the data collection or the analyses, the writing of the article, or in the decision to submit it for publication.

Informed consent Written informed consent was obtained from all subjects before the study, as all patients (or their caregivers/legally authorized representatives) included in the CULA North Oslo registry signed a written consent for usage of their data in future research studies.

Ethical approval INS applied for ethical approval from The Regional Health Trust's Ethical Committee, and the committee replied 21 December 2020 that the project was considered as a quality study and that ethical approval was not required.
Oslo University Hospital's Data Protection Officer approved the study 8 January 2021 (ID 21/00370). This study was completed in accordance with the Helsinki Declaration as revised in 2013.

ORCID iD Ida Neergård Sletten https://orcid.org/0000-0002-8884-3859

Supplemental material Supplemental material for this article is available online.

References

Baas M, Zwanenburg PR, Hovius SER, van Nieuwenhoven CA. Documenting combined congenital upper limb anomalies using the Oberg, Manske, and Tonkin classification: implications for epidemiological research and outcome comparisons. J Hand Surg Am. 2018, 43: 869.e1–e11.

Bae DS, Canizares MF, Miller PE et al. Intraobserver and interobserver reliability of the Oberg-Manske-Tonkin (OMT) classification: establishing a registry on congenital upper limb differences. J Pediatr Orthop. 2018, 38: 69–74.

Ekblom AG, Laurell T, Arner M. Epidemiology of congenital upper limb anomalies in Stockholm, Sweden, 1997 to 2007: application of the Oberg, Manske, and Tonkin classification. J Hand Surg Am. 2014, 39: 237–48.

Ezaki M, Baek GH, Horii E, Hovius S. IFSSH scientific committee on congenital conditions. J Hand Surg Eur. 2014, 39: 676–8.

Goldfarb CA, Ezaki M, Wall LB, Lam WL, Oberg KC. The Oberg-Manske-Tonkin (OMT) classification of congenital upper extremities: update for 2020. J Hand Surg Am. 2020, 45: 542–7.

Goldfarb CA, Wall LB, Bohn DC, Moen P, Van Heest AE. Epidemiology of congenital upper limb anomalies in a Midwest United States population: an assessment using the Oberg, Manske, and Tonkin classification. J Hand Surg Am. 2015, 40: 127–32.e1–2.

Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. Tutor Quant Methods Psychol. 2012, 8: 23–34.

Havenhill TG, Manske PR, Patel A, Goldfarb CA. Type 0 ulnar longitudinal deficiency. J Hand Surg Am. 2005, 30: 1288–93.

Huhnstock S, Svenningsen S, Merckoll E, Catterall A, Terjesen T, Wiig O. Radiographic classifications in Perthes disease. Acta Orthop. 2017, 88: 522–9.

Hülsemann W, Mann M, van Nieuwenhoven C, Sletten IN, Winge M. The European perspective on the Oberg-Manske-Tonkin classification: challenges for implementation, databases and registries. J Hand Surg Eur. 2020, 45: 1112–3.

James MA, McCarroll HR, Manske PR. The spectrum of radial longitudinal deficiency: a modified classification. J Hand Surg Am. 1999, 24: 1145–55.

Kallemeier PM, Manske PR, Davis B, Goldfarb CA. An assessment of the relationship between congenital transverse deficiency of the forearm and symbrachydactyly. J Hand Surg Am. 2007, 32: 1408–12.

Lam WL. App Store: Computer program. 2019a. Available at: https://apps.Apple.Com/gb/app/omt-medical-reference/id1465481577 (accessed 1 September 2020).

Lam WL. Google Playstore: Computer program. 2019b. Available at: https://play.Google.Com/store/apps/details?Id=com.Omt.Omtmedicalreference&hl=en_us (accessed 1 September 2020).

Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977a, 33: 159–74.

Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics. 1977b, 33: 363–74.

Lucas NP, Macaskill P, Irwig L, Bogduk N. The development of a quality appraisal tool for studies of diagnostic reliability (QAREL). J Clin Epidemiol. 2010, 63: 854–61.

Manske PR, McCarroll HR, James M. Type IIIa hypoplastic thumb. J Hand Surg Am. 1995, 20: 246–53.

Manske PR, Oberg KC. Classification and developmental biology of congenital anomalies of the hand and upper extremity. J Bone Joint Surg Am. 2009, 91(Suppl 4): 3–18.

McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012, 22: 276–82.

Oberg KC, Feenstra JM, Manske PR, Tonkin MA. Developmental biology and classification of congenital anomalies of the hand and upper extremity. J Hand Surg Am. 2010, 35: 2066–76.

Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther. 2005, 85: 257–68.

Temtamy SA, McKusick VA. The genetics of hand malformations. Birth Defects Orig Artic Ser. 1978, 14: i–xviii, 1–619.

Tonkin MA. Description of congenital hand anomalies: a personal view. J Hand Surg Br. 2006, 31: 489–97.

Tonkin MA, Oberg KC. The OMT classification of congenital anomalies of the hand and upper limb. Hand Surg. 2015, 20: 336–42.

Tonkin MA, Tolerton SK, Quick TJ et al. Classification of congenital anomalies of the hand and upper limb: development and assessment of a new system. J Hand Surg Am. 2013, 38: 1845–53.

Uzun H, Özdemir FDM, Üstün GG, Sakarya AH, Bitik O, Aksu AE. Oberg-Manske-Tonkin classification of congenital upper extremity anomalies: the first report from Turkey. Ann Plast Surg. 2020, 85: 245–50.