Plastic and Reconstructive Surgery Advance Online Article

DOI: 10.1097/PRS.0000000000009853

Quantifying the Severity of Metopic Craniosynostosis Using Unsupervised Machine Learning

Erin E Anstadt, MD[1]; Wenzheng Tao[2]; Ejay Guo[2]; Lucas Dvoracek[1], MD; Madeleine K Bruce, BA[3]; Philip J Grosse[4]; Li Wang[4]; Ladislav Kavan, PhD[2]; Ross Whitaker, PhD[2]; Jesse A Goldstein, MD[3]

1. University of Pittsburgh Medical Center, Department of Plastic Surgery, Pittsburgh, PA.
2. School of Computing, University of Utah; Salt Lake City, UT.
3. Department of Plastic Surgery, UPMC Children's Hospital, University of Pittsburgh Medical Center; Pittsburgh, PA.
4. Clinical and Translational Science Institute, University of Pittsburgh, Pittsburgh, PA.

Corresponding Author: Jesse A. Goldstein, MD Associate Professor, University of Pittsburgh Department of Plastic Surgery Children's Hospital of Pittsburgh One Children's Hospital Drive, 4401 Penn Avenue, Floor 3, #3533; Pittsburgh PA 15224 Telephone: (412) 692-8650 Fax: 412-692-8614 E-mail: jesse.goldstein@chp.edu

Financial Disclosures: This project was funded through a grant from the National Institutes of Biomedical Imaging and Bioengineering through Grant Number R21 EB026061, statistical analysis was supported in part by the National Institutes of Health through Grant Number UL1-TR-001857, and ShapeWorks was supported in part by the National Institutes of Health through grant numbers NIBIB-U24EB029011, NIAMS-R01AR076120, NHLBI-R01HL135568, NIBIB-R01EB016701, and NIGMS-P41GM103545.

Short Running Head Severity of Metopic Craniosynostosis

Abstract:

Background: Quantifying the severity of head shape deformity and establishing a threshold for operative intervention remains challenging in patients with Metopic Craniosynostosis (MCS). This study combines 3D skull shape analysis with an unsupervised machine-learning algorithm to generate a quantitative shape severity score (CMD) and provide an operative threshold score.

Methods: Head computed tomography (CT) scans from subjects with MCS and normal controls (age 5-15 months) were used for objective 3D shape analysis using ShapeWorks software and in a survey for craniofacial surgeons to rate head-shape deformity and report whether they would offer surgical correction based on head shape alone. An unsupervised machine-learning algorithm was developed to quantify the degree of shape abnormality of MCS skulls compared to controls.

Results: 124 CTs were used to develop the model; 50 (24% MCS, 76% controls) were rated by 36 craniofacial surgeons, with an average of 20.8 ratings per skull. The interrater reliability was high (ICC=0.988). The algorithm performed accurately and correlates closely with the surgeons assigned severity ratings (Spearman's Correlation coefficient r=0.817). The median CMD for affected skulls was 155.0 (IQR 136.4-194.6, maximum 231.3). Skulls with ratings ≥150.2 were highly likely to be offered surgery by the experts in this study.

Conclusions: This study describes a novel metric to quantify the head shape deformity associated with metopic craniosynostosis and contextualizes the results using clinical assessments of head shapes by craniofacial experts. This metric may be useful in supporting clinical decision making around operative intervention as well as in describing outcomes and comparing patient population across centers.

Introduction:

Metopic craniosynostosis (MC) refers to the premature fusion of the metopic cranial suture resulting in trigonocephaly, characterized by a keel-shaped forehead, suture ridging, orbital hypotelorism, retrusion and upsloping of the lateral supraorbital rim, and bitemporal narrowing.[1] While severe trigonocephaly is pathognomonic for MC, mild to moderate phenotypes are more difficult to diagnose and classify by severity. There is currently no standard for classification of severity, and variability in management protocols exists. Surgical intervention for mild to moderate phenotypes remains controversial as objective methods to delineate which patients require surgery are lacking.[1]

While operative intervention effectively preserves normal neurocognition, studies have demonstrated a worsening aesthetic results with longer follow-up.[2] Evidence is lacking regarding the role that pre-operative severity may play in this phenomenon due in part to the relative lack of objective methods to quantify the head shape deformity in MC.

Our group previously used a supervised machine learning algorithm to combine statistical head shape information with expert ratings to generate an objective measure of head shape deformity associated with MC.[3] It proved to correlate with expert ratings better than previously described methods, including interfrontal angle assessments as described by Kellogg et al.[4] This is attributed to the comprehensive nature of the three-dimensional shape analysis this study uses, whereas other techniques are often limited to two-dimensional measurements that do not account for subtle contour changes seen throughout the skull that may be perceived by surgeons on exam. The present study expands upon this model by utilizing an unsupervised machine learning algorithm that is not trained on severity data from experts, but rather uses only the known clinical diagnosis and raw shape data extracted from normal head computed tomography (CT)

images. The algorithm thereby generates an objective and quantitative measure, the Cranial Morphology Deviation (CMD), without the potential biases of clinician judgement or experience. The CMD represents the extent to which a patient's head shape deviates from the normal population in a way that allows visualization of deviations not strictly due to MC. We present the unsupervised algorithm's results and compare them to head shape severity ratings obtained from craniofacial surgeons. Moreover, we describe an operative threshold severity score that represents the degree of deformity at which the majority of surveyed surgeons offer surgical intervention. Ultimately, this study provides a mechanism to reduce subjectivity around the severity of MC, provides a potential standardized platform for communication and research between surgeons, and offers further insight into indications for operating on these patients.

Methods:

Head CT scans from subjects with MC and controls ages 5 to 15 months old seen at a tertiary children's hospital between 2002 and 2016 were used for three-dimensional (3D) skull shape modeling and expert ratings. All images were obtained using a standard low-dose, fine cut (0.25mm) protocol. Patients with MC were diagnosed by a board-certified craniofacial plastic surgeon using imaging as well as history and physical exam. Control subjects included age-matched patients who presented for trauma who demonstrated no abnormalities on CT head imaging and had a range of normal head-shapes.

Shape modeling and analysis: an unsupervised machine learning algorithm

CT scan processing and preparation for analysis using ShapeWorks software (University of Utah, Salt Lake City, UT) was performed as previously described.[3,5] Normal skull shape variations were quantified with a multivariate Gaussian distribution over a dense set of automatically-placed correspondence points. The controls were used to estimate the parameters

of the "normal" skull shape distribution against which MC-affected subjects were compared. The CMD was computed for each skull using Mahalanobis distance (negative log-likelihood), which quantifies the degree to which that skull shape differs from normal parameters.[6]

A randomly-selected set of normal control skulls was used to train the Gaussian model. The remaining normal skulls, as well as all of the metopic skulls, were reserved for testing the algorithm and calculating objective CMD values. The same test skulls were also used in a survey given to clinical experts to facilitate CMD comparison to perceptions of clinical severity.

CranioRate™ survey design

A web-based survey, CranioRate™ (https://www.craniorate.org/), was created and distributed to craniofacial surgeons. Eligible respondents included both craniofacial plastic surgeons and pediatric neurosurgeons. These clinical experts were recruited at regional, national, and international craniofacial surgery conferences, and contacted by email via the American Society of Craniofacial Surgery membership network.

Experts' demographic information and their head shape severity ratings were collected. Raters could rotate images of the 3D reconstructed CT scans for comprehensive visualization. Raters assessed the severity of head shape deformity for each skull on a 5-point Likert scale (0- normal to 4- most severe). They were also asked to decide if they would offer operative intervention based on skull shape alone. Initially, raters assessed all 50 skulls (60% of which had MC, 40% were control subjects) included in the study in a randomized order. The survey was later adjusted such that each expert rated a randomly-selected 20-skull subset from the total pool of 50 skulls to obtain input from more raters. Of note, experts could complete the survey in multiple sessions online to avoid environment factors or time constraints impacting results.

Statistical analysis

Statistical analysis was performed using R (R Core Team, Vienna, Austria). Descriptive statistics were calculated. The intraclass correlation coefficient (ICC) was calculated to assess agreement on skull ratings between experts. A consensus severity rating for each skull based on aggregation of expert ratings was estimated using a maximum likelihood estimation (MLE) approach. A consensus decision on the need for surgical correction was determined for each skull based on majority response (>50%) from the raters.

Spearman correlations were used to compare the CMD values to the consensus severity ratings from the expert survey. A Kruskal-Wallis test and Bonferroni adjusted p-values for the multiple comparisons was used to compare differences in CMD scores between severity groups.

To validate the CMD as a tool to discern MC from normal shapes, point-biserial correlations were used to compare CMD results to the known diagnosis (control versus affected skull). A pooled logistic regression cross-validation model was then used to analyze the accuracy of the CMD score in predicting skull's status as normal versus affected. A receiver operator characteristic curve (ROC) was created and the area under the curve (AUC) was calculated to evaluate the goodness of fit of the model.

Finally, a leave-one-out cross validation model was performed to determine the accuracy of using CMD to predict raters' decision to operate.[3]

Significance was set at alpha = 0.05 for all statistical analysis.

RESULTS:

Head CT scans from 124 patients were used: thirty (24%) of these patients had MC while the remaining 94 (76%) were controls. Seventy-four randomly-selected normal skulls (78%) were

used for training the machine learning algorithm; the remaining 20 normal skulls, and all 30 craniosynostotic skulls, were reserved for testing and the survey.

CranioRate™ survey results:

36 surgeons completed the survey; sixteen (44%) rated all 50 skulls and 20 (56%) rated a randomly-selected subset of 20 skulls. The mean number of raters per skull was 20.8 and the total number of ratings obtained was 1,200.

Rater demographics: Forty percent of raters have been in practice for over 15 years (Table 1). The majority (85.7%) were plastic surgeons. Most raters (n=23, 64%) treat approximately 10 to 50 MC patients yearly, and most perform <15 surgeries per year for MC (n=29, 81%). Academic practice settings were most common (n=26, 72%), followed by a combined academic/private practice (n=7, 19%).

There was a high level of agreement between raters' severity scores (ICC= 0.988). According to consensus severity scores, 25 skulls had mild deformity, 9 were moderate, and 16 were severely deformed. No skulls had a consensus severity rating of "most severe."

Severity rating patterns:

Normal skulls showed highly consistent severity ratings by the surgeons. Less agreement was observed in severity ratings of affected skulls compared to controls, particularly for affected skulls with "moderate" severity ratings. Consensus severity ratings correlated to known skull status (metopic versus control), indicating surgeons rated affected skulls higher than controls (point-biserial correlation 0.745 ($p<0.0001$).

Shape Modeling and Machine Learning Algorithm results:

Cranial Morphology Deviation (CMD) scores:

The median CMD for control skulls was 93.5 (IQR 88.6-100.7, maximum 118.7). The median CMD for affected skulls was 155.0 (IQR 136.4-194.6, maximum 231.3). Affected skulls tend to have higher CMD scores than normal skulls; the point-biserial correlation for CMD with known skull status (metopic versus control) was 0.703 ($p<0.0001$).

The cross-validation analysis showed that the CMD can predict true skull status (control versus affected) with high level of accuracy (0.881, standard deviation 0.059) and consistency. Using this model, skulls with an CMD score of 111.9 or greater are more likely to be truly pathologic (sensitivity 90%, specificity 90%, AUC=0.9383).

CMD scores tend to increase as consensus severity ratings increase according to the raters. Spearman correlation between CMD and consensus severity scores was r=0.817, implying a strong correlation between this algorithm and the surgeons' clinical assessment of severity (Table 2).

Further descriptive analysis was performed to understand the distribution of CMD scores within each category of consensus severity ratings by surgeons. There was a significant difference in CMD scores between severity groups, $\chi^2(2) = 37.165$, $p < .001$. The mild severity group was found to have significantly lower CMD scores than the moderate severity group ($p = .013$) and the most severe group ($p < .001$). The moderate severity group CMD scores did not differ significantly from the most severe group ($p = .150$). Box plots of CMD distributions within each severity rating category demonstrate little overlap, suggesting the CMD scores reflect true differences in clinical severity (Figure 1) as detected by the raters. Furthermore, these plots show that CMD varies more among skulls with more severe deformity according to raters. A clinical rating of "severe" actually represents a heterogenous group of skulls with a spectrum of severity.

Decision-to-operate results:

Of the 1,200 ratings given, raters gave 495 (41.3%) "Yes" responses when asked whether they offer corrective surgery based on skull shape alone (Table 3).

This data suggests a strong level of agreement among raters' decision to operate when they perceived severity to be moderate or worse (rating of ≥2). When the decision-to-operate data was compared to the consensus scores, the Spearman correlation was 0.975, implying a strong positive linear relationship between level of perceived severity and proportion of raters who would offer corrective surgery (Table 2).

The Spearman correlation between CMD and the proportion of raters deciding to operate was 0.849 ($p<0.0001$) (Table 2). All skulls that had a majority of raters (>50%) electing to offer corrective surgery had an CMD ≥111.9 (Figure 2). There was very high agreement between experts on the need for surgery for patients with CMD >150.2, with nearly all of these skulls having >90% consensus on electing to offer surgery. No control skulls had ≥25% of raters opting to operate.

Finally, a Leave One Out cross-validation analysis evaluated accuracy of the CMD algorithm in predicting the decision to operate of each expert rater for an unseen case. It can be used to calculate the likelihood of a skull receiving a positive decision to operate by the majority of expert raters based on its CMD. This prediction model was compared to the actual record of each rater electing to offer surgery using the Mean Squared Error (MSE). MSE was 0.158, representing the accuracy of the prediction on decision to operate being 84.2%.

DISCUSSION

Assessing the severity and prognosis of MC remains difficult despite attempts to standardize clinical evaluations. Studies in this arena are largely based on anthropomorphic measurements of trigonocephaly.[7–10] However, these efforts are limited in their ability to discriminate mild and moderate MC from normal variants, and are likely impaired by reliance upon discrete, two-dimensional variables to describe this complex 3D deformity.

In the present study, we describe our updated model for assessing MC severity, which combines comprehensive, 3D shape analysis with a novel unsupervised machine learning algorithm to provide a severity score, the CMD, which is quantitative and independent from individual clinician bias.[3] Additionally, we present an analysis of MC severity ratings from 36 surgeons to provide context for our CMD data and identify potential biases in judging severity. Moreover, this is the first study to suggest a threshold for surgical intervention that is based on evaluation of 3D skull morphology in conjunction with consensus opinions from 36 surgeon experts.

This analysis establishes several important traits of our metric, CMD: it is higher for affected skulls than for controls, it correlates with observed clinical severity, and it can reliably predict true skull status (normal versus affected).

The expert rater cohort included in this study represents a mix of practitioners in various stages of their career, from different subspecialties and practice settings, and with different patient populations. Overall, the majority of raters appropriately rated the normal, control patients in 90% of cases. Within the "gray zone" of skulls with clinical assessments of "moderate" severity, the rank-ordering of the skulls varied slightly when comparing consensus scores from raters' to the objective CMD values. The availability of an objective CMD data point could be especially useful to inform clinical decision-making in these moderate cases. Furthermore, the skulls

assigned “severe” ratings by experts had a wider range of CMD values than skulls with mild deformities, indicating that the severe group represents a heterogenous population. Indeed, the clinical rating of “severe” may be inadequate to describe the variability of head shapes in this group, highlighting the need for an unbiased continuous metric to provide more granularity within severity subgroups than may be possible or reasonable to delineate by clinical judgement alone. For less experienced surgeons, this objective measure could provide an invaluable reference to guide decision-making. Importantly, the CMD is not meant to replace the nuanced process of physicians’ decision-making as a whole. This study indicates that patients with a certain degree of head shape difference are more likely to be offered surgery by the majority of experts polled in this study. This data can be directly used in patient/parent counseling. It is meant to serve as an assist to surgeons and families that provides a greater understanding of how a given patient’s condition fits into the spectrum of shape deformity. While no algorithm or consensus opinion can completely replace physician judgment for individual patient cases, this algorithm offers an objective means by which physicians can approach their assessment and subsequent management of patients.

This study is the first to compare severity ratings with operative decision making. Expert raters were consistent with their responses and confirmed the notion that surgeons are less likely to operate on milder cases. At least 90% of raters felt the “severely” rated skulls (correlating to an CMD of 188.4) warranted operative intervention. The cross-validation analysis demonstrates that for a new skull introduced to this model, we can use the CMD value and expert rater thresholds to predict these experts’ decision to operate with a high level of accuracy. This may be particularly valuable for moderately-affected skulls with an uncertain need for surgery. Using

this CMD data also allows surgeons to better describe clinical outcomes relative to pre-operative head shape.

While some skulls were rated as "most severe" by individual raters, no skulls had a consensus score of "most severe." We acknowledge that this may limit the generalizability of these results and further study is needed. That said, patients with the most severe deformity often have clearer indications for surgery. Additionally, no patients with metopic ridge were included in this cohort, as our institutional protocol is to follow these patients clinically rather than obtaining a CT scan. Further studies are underway in collaboration with other institutions to expand the capabilities of our model. Given the nature of the machine learning algorithm, we acknowledge that the CMD reflects abnormalities in head shape that are inherently not specific to MC alone. Thus, normal patients with subtle skull shape changes for other reasons, such as deformational plagiocephaly, can also yield higher CMD scores than may be expected in normal patients. Similarly, a patient who is rated clinically as having mild metopic CS may yield a high CMD because the objective skull shape is statistically different from normal despite that shape not being clinically significant to the physician observers. Furthermore, we acknowledge that 36 experts alone cannot perfectly represent the diversity of perspectives that exist within the body of craniofacial surgeons globally. We attempted to mitigate this by recruiting raters by various means and offering raters the opportunity to complete the survey on their own time where environmental constraints are less likely to impact results. We acknowledge that individuals may inherently differ in the amount of time they are willing to commit to completing a survey and data may be affected accordingly. That said, we feel this data provides a substantial and deeper understanding of many surgeons' approach to the management of these patients.

While previous studies have explored the application of statistical shape analysis and modeling in craniofacial surgery, there have not been mechanisms available to apply this technology clinically.[11] The ideal model is reproducible and leads to meaningful stratification of severity that corresponds to a likelihood of benefitting from operative intervention. We are currently using our data and algorithm to develop an online portal (https://www.craniorate.org/) to which clinicians can upload deidentified CT scans for analysis by our algorithm. Individual patient CMD scores will be calculated and compared to other patients with MC. Given the accuracy with which our statistical model can predict the surgeons' opinion regarding need for surgical intervention for a patient with a particular CMD score, this tool will also provide insight into whether a majority of surgeons would choose to operate. Ultimately, we hope to expand this algorithm to include other craniosynostosis types and develop the algorithm's ability to distinguish them as well as validate our model on 3D surface topography photos/scan so as to allow stratification even in the absence of CT data.

CONCLUSIONS

This study describes a novel metric to quantify the head shape deformity associated with metopic craniosynostosis and contextualizes the results using clinical assessments of head shapes by craniofacial experts. While its results are largely consistent with clinical assessments, it provides increased granularity across the spectrum of severity. This metric may be useful in supporting clinical decision making around operative intervention as well as in describing outcomes and comparing patient population across centers.

References:

1. Jaskolka MS. Current Controversies in Metopic Suture Craniosynostosis. *Oral Maxillofac Surg Clin North Am*. 2017;29(4):447-463. doi:10.1016/j.coms.2017.07.003
2. Wes AM, Paliga JT, Goldstein JA, Whitaker LA, Bartlett SP, Taylor JA. An evaluation of complications, revisions, and long-term aesthetic outcomes in nonsyndromic metopic craniosynostosis. *Plast Reconstr Surg*. 2014;133(6):1453-1464. doi:10.1097/PRS.0000000000000223
3. Bhalodia R, Dvoracek LA, Ayyash AM, Kavan L, Whitaker R, Goldstein JA. Quantifying the Severity of Metopic Craniosynostosis. *J Craniofac Surg*. 2020;00(00):1. doi:10.1097/scs.0000000000006215
4. Kellogg R, Allori AC, Rogers GF, Marcus JR. Interfrontal Angle for Characterization of Trigonocephaly. *J Craniofac Surg*. 2012;23(3):799-804. doi:10.1097/SCS.0b013e3182518ad2
5. Cates J, Elhabian S, Whitaker R. ShapeWorks. In: *Statistical Shape and Deformation Analysis*. Elsevier; 2017:257-298. doi:10.1016/B978-0-12-810493-4.00012-2
6. De Maesschalck, Roy; Jouan-Rimbaud, Delphine; Massart DL. The mahalanobis distance. 50.1 (2000): 1-18. *Chemom Intell Lab Syst*. 2000;50(1):1-18.
7. Anolik RA, Allori AC, Pourtaheri N, Rogers GF, Marcus JR. Objective Assessment of the Interfrontal Angle for Severity Grading and Operative Decision-Making in Metopic Synostosis. *Plast Reconstr Surg*. 2016;137(5):1548-1555. doi:10.1097/PRS.0000000000002052
8. Farber SJ, Nguyen DC, Skolnick GB, Naidoo SD, Smyth MD, Patel KB. Anthropometric Outcome Measures in Patients With Metopic Craniosynostosis. *J Craniofac Surg*.

2017;28(3):713-716. doi:10.1097/SCS.0000000000003495

9. Beckett JS, Chadha P, Persing JA, Steinbacher DM. Classification of Trigonocephaly in Metopic Synostosis. *Plast Reconstr Surg*. 2012;130(3):442e-447e. doi:10.1097/PRS.0b013e31825dc244

10. Wang JY, Dorafshar AH, Liu A, Groves ML, Ahn ES. The metopic index: an anthropometric index for the quantitative assessment of trigonocephaly from metopic synostosis. *J Neurosurg Pediatr*. 2016;18(3):275-280. doi:10.3171/2016.2.PEDS15524

11. Mak ML, Al-Shaqsi SZ, Phillips J. Prevalence of Machine Learning in Craniofacial Surgery. *J Craniofac Surg*. 2020;00(00):1. doi:10.1097/scs.0000000000006234

Figure Legends:

Figure 1. Box plot showing CMD distribution within each category of severity ratings (mild, moderate, severe) as determined by MLE consensus score from clinical experts. Mean CMD score ± standard deviation is reported below each severity group.

Figure 2. Decision to operate compared to CMD. Normal skulls (green) are clustered around a lower CMD, with very few raters electing to offering operation.

Tables

Table 1. Rater demographic information.

Year in Practice	N (%)
0 to 3	5 (14%)
3 to 5	4 (11%)
5 to 10	4 (11%)
10 to 15	7 (19%)
Greater 15	14 (39%)
Unknown	2 (6%)
Area of Practice	**N (%)**
Neurosurgery	5 (14%)
Plastic Surgery	30 (83%)
Unknown	1 (3%)
Current Surgical Status	**N (%)**
In Practice	33 (92%)
In Surgical Training	3 (8%)
Job Title	**N (%)**
Attending Surgeon	33 (92%)
Fellow	2 (2%)
Unknown	1 (1%)
Number of Annual Patients	**N (%)**
0 to 10	8 (22%)

10 to 50	23 (64%)
50 to 100	4 (11%)
Unknown	1 (3%)
Practice Setting	**N (%)**
Academic Practice	26 (72%)
Private Practice	3 (8%)
Combined	7 (19%)
Surgery Frequency (cases per year)	**N (%)**
0 to 5	9 (25%)
5 to 10	14 (39%)
10 to 15	6 (17%)
15 to 20	3 (8%)
20 to 50	3 (8%)
Unknown	1 (3%)

Table 2. Spearman and point-biserial correlations between severity, CMD, true/known skull status and decision to operate. The point-biserial correlation coefficient measures the strength of the association between a continuous variable and a dichotomous variable with 1 being the strongest possible correlation.

	CMD	True Skull Status	Decision to Operate	p-value
Consensus Severity	0.817	0.76	0.975	<0.001
CMD	-	0.703	0.849	<0.001

Table 3. Decision to operate by severity rating.

Rating	Number of "Yes" Responses	Number of Ratings	Percentage of "Yes" Responses
0	5	451	1.1%
1	44	287	15.3%
2	220	236	93.2%
3	180	180	100.0%
4	46	46	100.0%

Figure 1

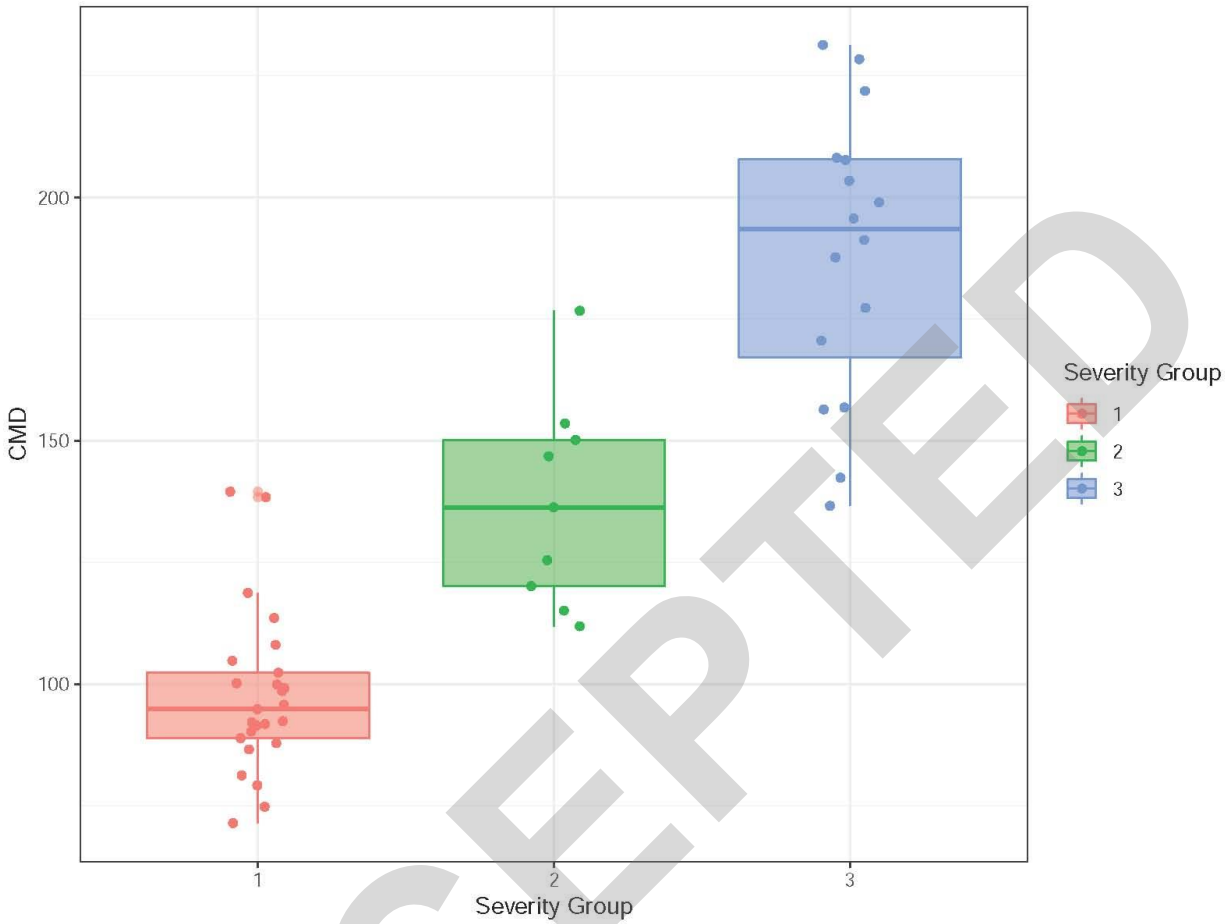

Figure 2