Reliability of a standardized protocol to calculate cross-sectional chest area and severity indices to evaluate pectus excavatum

M. Louise Lawsona,b, Myra Barnes-Eleya, Bonnie L. Burkea,c, Karen Mitchella, Michael E. Katzd, Christopher L. Dorye, Stephen F. Millerf, Donald Nussg, Daniel P. Croitorug, Michael J. Goretskyg, Robert E. Kelly Jrg,*

aCenter for Pediatric Research, Children’s Hospital of The King’s Daughters and Eastern Virginia Medical School, Norfolk, VA, USA
bDepartment of Surgery, Center for Epidemiology and Biostatistics, Cincinnati Children’s Hospital and Medical Center and University of Cincinnati, Cincinnati, OH, USA
cDepartment of Pediatrics, Epidemiology and Biometry Core, Eastern Virginia Medical School, Norfolk, VA, USA
dDepartment of Radiology, Imaging Consultants of South Florida, West Boca Raton Medical Center, West Boca Raton, FL, USA
eDepartment of Radiology, Children’s Hospital of The King’s Daughters and Eastern Virginia Medical School, Norfolk, VA, USA
fThe Hospital for Sick Children, Toronto, Canada
gPediatric Surgery, Department of Surgery, Children’s Hospital of The King’s Daughters, Eastern Virginia Medical School, Norfolk, VA, USA

Abstract

Purpose: In evaluating the impact of surgical repair of pectus excavatum, the Haller index developed for preoperative decision-making purposes may be inadequate to quantify postoperative changes in shape of the chest. Individual patients may also have chest characteristics that impact the success of repair, many of which would be unlikely to be measured by the Haller index alone. We have developed a protocol that measures the cross-sectional chest area and the asymmetry index along with the Haller index to more completely quantify the nature of the deformity. The purpose of this study was to determine the reliability of this protocol in the interpretation of chest computed tomography images from multiple sites. The protocol was developed as part of a multicenter study of clinical outcomes after surgical repair of pectus excavatum.

Methods: Two radiologists independently selected 5 images from each of 32 computed tomography scans from multicenter study participants according to the protocol. A digitizer was used to measure the diameters and cross-sectional areas of the images selected; these results were used to calculate the Haller and asymmetry indices. The protocol was tested for intradigitizer and interradiologist reliability. Using the Haller and asymmetry indices, we also assessed agreement between radiologists classifying patients as abnormal.

Results: Agreement was uniformly high for all comparisons (all Lin’s concordance coefficients >0.99 and all Cohen’s k’s >0.85, all agreement on classification of patients >95%) indicating almost perfect agreement. Disagreement on classification of patients using the Haller and asymmetry index was at the cut points for abnormality.
Pectus excavatum is the most common congenital chest wall deformity, affecting about 23 per 10,000 births [1]. Whether surgical repair of the deformity is medically necessary depends on the impact that the pectus deformity has on physiology, which is partially dependent upon the depth of the deformity. The Haller index was developed to provide an objective measure of the depth of the deformity for determining surgical eligibility [2,3]. In addition to the Haller index, we have recently begun measuring the asymmetry index because we have noted in our surgical practice that more asymmetrical defects can be significantly harder to repair and can have less favorable outcomes.

In the several hundred patients we have seen with pectus excavatum, we have noticed considerable variability among medical practitioners in determining the Haller and asymmetry indices, depending on how the images are chosen and how measurements are taken from the chosen images.

As a part of a multicenter study in which impact of surgical repair on the shape of the chest is one of the primary outcome measurements, we developed a standard protocol to help control for variation in deriving these indices. Our protocol defines the selection of images that are then digitized to allow the assessment of the depth of the depression toward the spine, its right or left asymmetry, and the cephalocaudal extent of the depression. The protocol also provides measurements of cross-sectional areas of the chest at different points, which we believe may be a better measurement of surgical outcome than the indices. Patients with deep chests, as in “barrel-shaped” chests, may have deceptively low Haller indices. In addition, repair could lead to an improvement in symmetry and depth of the depression without significantly improving the chest volume. The cross-sectional area will allow us to better approximate the extent of improvement in the total chest volume after repair.

The goal of this study was to determine the interobserver agreement of our measurement protocol and the precision of the digitizer measurements when applied to computed tomography (CT) scans.

1. Methods

1.1. Subjects

All participants in the multicenter study signed an informed consent that was in compliance with the human experimental guidelines of the United States Department of Health and Human Services and of the Eastern Virginia Medical School Institutional Investigation Review Board. The 32 participants were predominately male (91%), and the average age was 13.2 years (range, 6.1-19.7 years). Because of the timing of this reliability study early in the multicenter study, all 32 patients were from the Children’s Hospital of the King’s Daughters (CHKD) surgical population. The CT window was bone in 25 of the patients, lung in 3 patients, soft tissue in 1 patient, and a combination of soft tissue and bone in 3 patients. The CT scans were taken at 20 sites around the United States, with 13 of them being taken at CHKD. There was no difference in results between scans taken at CHKD and at other sites.

1.2. Study design

In the multicenter study, the surgical outcomes of interest included changes in anatomy of the chest, pulmonary functioning, and quality of life. Our use of pulmonary function and quality-of-life measures in these patients has been documented elsewhere [4,5]. We used CT scans of the chest with the interpretation protocol described below to measure the baseline anatomical impact of the deformity and for measuring the anatomical impact of the repair. An essential part of surgical planning, preoperative CT scans were taken according to the protocol developed by one of the radiologists (MEK) associated with the study. The purpose of the detailed protocol described here is to standardize the interpretation of the CT scans from multiple sites around the country. The study design calls for the CT scans to be sent to a central location and examined independently by 2 radiologists. If the 2 radiologists come to conclusions that are different, a third radiologist is available to serve as arbiter of the final decision. Thus, we needed to determine if the proposed method could be applied consistently by different interpreters and would yield reliable, reproducible results. One objective of the present reliability study was to determine what level of difference in interpretation could be considered within the realm of chance and what level of difference should result in arbitration by the third radiologist in the multicenter study.

For this reliability study, the 2 radiologists (CLD, SFM) who are serving as the primary study radiologists in the larger study examined 32 CT scans from the first 6 months of the data collection. Each radiologist independently used the protocol described below to determine which CT images should be used to calculate the cross-sectional area and Haller and asymmetry indices. The images selected captured the deformity at various points, providing a view of the total...
deformity as opposed to the traditional measurement of 1 ratio at a single level. The study coordinator then used a digitizer to scan each of the images twice, allowing us to assess both intradigitizer (comparing the first and second results of a single image from the digitizer) and interradiologist (comparing the first results from the digitizer on the different images chosen by the different radiologists) agreement on the raw measurements and on the 2 indices of severity. Because the radiologists sometimes selected the same images for a given CT, a specific image might have been scanned up to 4 times, although not consecutively.

1.3. Computed tomography scan protocol

1. The CT scan should be at least from above the manubrium to below the xiphoid on a patient who is supine, the long axis of the midthorax being aligned with the long axis of the table. Ideally, both the pre- and post-CT are done with inspiratory breath holding, using 5-mm scans.

2. Bone windows (3000/500) were preferred for measurements when possible. Lung (1500/−500) and soft tissue (400/40) were used if bone windows were not available.

3. Measurements were made at 5 levels on each examination (see Fig. 1):
   a. position 1: the level of the sternomanubrial junction (anterior second rib ends);
   b. position 5: the level of the tip of the xiphoid;
   c. position 4: the level of the end of the body of the sternum;
   d. positions 2 and 3: divide the distance between positions 1 and 4 by 3. Position 2 is one third of the way between positions 1 and 4, and position 3 is two thirds of the way between positions 1 and 4 (calculated by the digitizer technician).

4. Measurements were made in the intrathoracic area of each slice, including the lungs, heart, and mediastinum plus any pleural thickening or fluid and excluding the spine, ribs, and chest wall musculature (see Fig. 2).

5. Measurements were made with a digitizer (see digitizer protocol) from the hard copy for each of the 5 selected levels:
   a. \( T \): maximum transverse internal thoracic diameter;
   b. \( A \): minimum AP diameter perpendicular to \( T \) between the coronal plane of the deepest part of the anterior chest wall and the coronal plane of the anterior aspect of the vertebral column;
   c. \( R \): maximum AP diameter on the right hemithorax, measure perpendicular to \( T \);
   d. \( L \): maximum AP diameter on the left hemithorax, measure perpendicular to \( T \); and
   e. cross-sectional area is calculated using the digitizer’s area function and tracing the inner circumference of the chest wall with the digitizer cursor.

After the measurements were made, the study coordinator used Excel to calculate the Haller index for each slice as \( T/A \). The asymmetry index was calculated for each slice as \( R/L \times 100 \). For the multicenter study and in clinical practice, a patient’s overall Haller index is defined as the largest of the 5 calculated; the asymmetry index is defined as the farthest from 100 of the 5 calculated. For this reliability study, we compared the calculated Haller and asymmetry indices between successive digitizer measurements and between radiologists for each slice selected, as well as assessing the reliability of the method and it would be used in clinical practice (ie, selecting the most extreme values of each index).

1.4. Digitizer protocol

Each radiologist selected images corresponding to positions 1, 4, and 5. The study coordinator then calculated
positions 2 and 3 and measured each of the specified data points 2 times for each image selected by the radiologists. The GTCO Surface-Lit AccuTab Digitizer was used to measure parameters from chest CT scans (detailed specifications available at http://www.gtc.com/productsurfacelitaccutab.htm#specs).

Measurements were dropped into a Microsoft Excel spreadsheet directly from the digitizer, which eliminated the possibility of human transcription errors.

With the CT scan secured to the digitizer surface, the scale and units were set and measured for each individual scan with the transducer. Transparent grid sheets were used to align measurements and ensure that lines were measured parallel or perpendicular, as appropriate, to the coronal plane. Intrathoracic area measurements were done by marking the start and stop point within the image.

For repeated digitizer scans used to measure intradigitizer reliability, the study coordinator used the digitizer to measure the chosen images in sequential order (eg, images 1-5 were scanned in order and then scanned again in the same order rather than image 1 being scanned twice consecutively).

### 1.5. Statistical analysis

The median was used as the measure of central tendency because many of the comparisons involved nonnormally distributed data. For normally distributed data, the mean and median are similar, so we chose to use the median throughout for consistency. Significance was set at the 95% confidence interval (95% CI) of the median, which we calculated using a bootstrap method [6].

Reliability of measurements made according to the protocol was assessed by examining agreement between the raw measurements, including cross-sectional area, and the Haller and asymmetry indices calculated by the study coordinator using the digitizer. Reliability of image selection according to the protocol was assessed by examining agreement between the radiologists on which image was chosen for each of the 5 positions. Agreement between the selections made by the radiologists (intraradiologist) and between the measurements made by the digitizer (intradigitizer) was analyzed using Lin’s concordance coefficient, which is more appropriate than Cohen’s $\kappa$ for data that are continuous and discrete. The concordance coefficient evaluates the agreement between 2 measures and ranges in value from $-1.0$ to $+1.0$. A concordance coefficient of at least 0.90 was needed to demonstrate the desired reliability of the method. The study was powered so that we could detect an average difference of at least 0.010 on raw scores between raters.

It was important not only to establish the protocol as a reliable way to select and measure CT images, but also to ensure that the protocol provided a reliable method for classifying patients for surgical eligibility. To assess the reliability of our protocol for classifying patients, we used the methods described by Kundel and Polanski [7]. Agreement for the categorical outcome “clinically normal” was assessed using Cohen’s $\kappa$. The recommended criteria for substantial agreement is indicated by a $\kappa$ between 0.61 and 0.80, with a $\kappa$ greater than 0.81 being almost perfect agreement. In this study, we declared that there was a difference in classification using the Haller index when the 2 radiologists disagreed in their placement of the patient above or below the 3.2 threshold, which is used as the cut point for eligibility for surgery by insurance companies and surgeons [8].

Although no cut point for surgical eligibility has been set for the asymmetry index, this index is a likely predictor of surgical outcome. Thus, our ability to measure this index accurately is important. Because the asymmetry index is a ratio of 2 sides of the pectus excavatum depression, values away from 100 indicate increasing asymmetry. Whether the index is greater than or less than 100% is merely a reflection of whether the right or the left side of the depression is deeper. We set artificial cut points, using the top and bottom 10th percentile in this population, for the asymmetry index to assess the potential for interrater agreement on classification of patients using this index. An asymmetry index outside the range of the 10th to 90th percentile indicates an asymmetric defect for the purposes of this analysis.

We determined the expected “real-world” differences between the radiologists on classification of patients by comparing the classifications of the most severe Haller and asymmetry indices as measured by each radiologist for a patient because this is the manner in which the indices are actually used in the clinical realm. To determine if the differences between radiologists would affect the interpretation of the cross-sectional area, we ranked the largest area chosen for each patient by each of the radiologists and compared the rankings between radiologists. This method of comparison is similar to what is done with a nonparametric statistical test of difference. We used it in this instance to provide an estimate of the actual impact of differences between the radiologists on the relationship between area measurements and used Lin’s coefficient as previously described for the appropriate statistical test of agreement. We also assessed a “worst-case scenario” by examining the impact on classification using the Haller indices of the largest observed difference between the radiologists among all of the 32 patients for each of the 5 CT images. For all analyses except the calculation of Lin’s concordance coefficients, which was done in Excel using output from PROC MEANS in SAS, SAS version 8.0 (Cary, NC) was used.

### 2. Results

The protocol used by the radiologists included detailed instructions on how to select the appropriate 5 images (see Methods). Each of the 32 scans had up to 6 pages of images that could be selected for calculation of pectus defect severity. Reliability between the 2 radiologists was entirely dependent on the images chosen and the impact of any
differences in the images on the raw measurements. The 2 radiologists chose within 1 image of each other most of the time (81%; n = 160) and were within 3 images 97% of the time. Lin’s concordance coefficient for the images chosen was consequently 0.99 (95% CI, 0.99-0.99). Because there was such strong agreement between the CT images chosen by the 2 radiologists, we can assume that any differences observed in the Haller and asymmetry indices were not primarily because of differences between the radiologists. Agreement between the raw values in the repeated calculations using the digitizer (n = 320 images each repeated twice) was equally high, with a concordance coefficient of 0.99 (95% CI, 0.99-0.99). This indicates that there was little variation because of either technical problems with the digitizer or to the technician in using the digitizer.

Table 1 shows the overall agreement on measurements for the severity indices. Intradigitizer and interrater agreement was extremely high, with all concordance coefficients greater than 0.99 (95% CI, 0.99-0.99) indicating near perfect agreement. Our results indicate that the expected variation between observers is no larger than 1.8% (the largest median percent change), and the variation that could be attributed to chance would be between 0.4 and 2.2 (see 95% CI of median percent change).

Comparing the classification of severity using the protocol, we see similar near-perfect agreement between observers for both the Haller and the asymmetry indices. There was agreement on categorization more than 95% of the time, with a $\kappa$ greater than 0.8 for all comparisons, indicating near perfect agreement (Table 2). The intradigitizer disagreements on the asymmetry index were because of 7 individuals with an asymmetry index close to the cut points of 86 (range, 85-88) or 106 (range, 105-108). Likewise, the intradigitizer disagreements on the Haller index represented 3 individuals with Haller indices close to 3.2 (3.19-3.23).

Table 2 shows the agreement in classifying pectus excavatum severity. Intradigitizer and interrater agreement was extremely high, with all Cohen’s $\kappa$ values greater than 0.85 (Table 2). To determine the potential impact of this “real-world” variation on clinical practice, we assessed the impact of observer differences on the maximum indices calculated by each radiologist, which is how indices should be assigned in clinical practice with this protocol. In this analysis (n = 32), we determined the agreement between the radiologists to be extremely high, with only 1 individual for whom the raters disagreed on the Haller index (no statistic possible because of zero cells) and 2 for whom the raters disagreed on the asymmetry index ($\kappa = 0.79$).

To determine agreement for the raters of cross-sectional area measurement, we sorted the patients based on the largest area determined by each radiologist. Each patient was given a rank as to their order of cross-sectional area. We compared rankings of the largest area for each patient between the 2 radiologists and saw that the 2 radiologists ranked the patients exactly the same way in their relative cross-sectional areas. Thus, the radiologists had perfect agreement on the relative area measurements, which resulted in the high Lin’s coefficient of 0.99 (Table 1).
Finally, to assess the impact of the maximum differences measured between radiologists, we determined the maximum difference in Haller indices calculated on each of the 5 cuts. Table 3 shows the results of this analysis, indicating that even the largest differences observed during the study had no impact on the classification of the patient using the Haller index. The smallest variation between the radiologists was seen at position 2, whereas a single large variation of more than 50% was seen at the xiphoid image. The next largest variation between radiologists at the xiphoid image (position 5) was 12%, again with no change in clinical interpretation.

### Table 3: Impact of maximum interrater differences at each CT image (n = 32)

<table>
<thead>
<tr>
<th>CT scan image position</th>
<th>Maximum absolute difference in Haller indices between radiologists</th>
<th>Lin's concordance coefficient</th>
<th>% Change of Haller interpretation</th>
<th>Change in interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.43</td>
<td>0.99</td>
<td>10</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>0.27</td>
<td>0.99</td>
<td>6</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>0.37</td>
<td>0.99</td>
<td>9</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>0.78</td>
<td>0.99</td>
<td>16</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>6.25</td>
<td>0.66</td>
<td>53</td>
<td>No</td>
</tr>
</tbody>
</table>

*a From the largest of the 2 Haller index values calculated for that image.*

3. Discussion

The 5-position standard protocol was proposed to alleviate potential biases and inconsistencies in data being collected from multiple centers with competing surgical treatments. Although the protocol is more extensive than just determining a single Haller index as a rough gauge of severity, it provides a tool for assessing both the need for surgery and the outcome of repair in any future quality monitoring program or to readily study any potential future modifications of surgical technique. To be considered an effective tool that fulfills its purpose, the proposed protocol must be shown to be reproducible by individual observers and to have solid interobserver agreement. In this study, we have demonstrated that our standardized protocol has excellent interrater reliability in calculating the cross-sectional area and the Haller and asymmetry indices and is robust to significant variation in observer choices.

In determining surgical necessity, often the Haller index is not calculated by a radiologist, but by a clinician. Periodically, we have seen Haller indices calculated by referring clinicians that are substantially different from those calculated by the surgeons participating in our multicenter study. This protocol provides a standardized method of deriving the Haller index, regardless of who makes the measurement. By using a standard calculation method, it is hoped that either a frontline clinician, a radiologist, or support staff could obtain the same results. For the purposes of the multicenter study, the use of the digitizer minimizes the likelihood of human error in the calculations. However, in clinical settings where a digitizer may not be available, all modern CT scanners and picture archive and communication system workstations have the capability to measure the prescribed area of interest.

One of the primary strengths of the protocol is its use of the cross-sectional area of the CT. This enables the clinician to assess the impact of surgery on the volume of the chest as a whole. Measuring the chest configuration as a whole can give a clearer picture of the extent of the deformity. For example, patients with barrel-shaped chests (narrow, oval shape) are often found to have a lower Haller index, which may not accurately represent the extent of the deformity. At this point, there are no recommendations for a minimum cross-sectional area of the chest that would result in clinical intervention, but such a recommendation would be useful now that we have demonstrated that the area can be reliably measured.

Another important strength of this study is that it provides a reliable means for measuring an asymmetry index. Patients with pectus excavatum are sometimes noted to have a degree of asymmetry in their chest, but measurement of the extent of asymmetry has not generally been done. Our experience has shown that the degree of asymmetry can affect the surgical outcome of pectus excavatum repair. Knowing the degree of asymmetry and its impact on surgical outcome would help in counseling patients as to what surgical outcome they could expect.

Although it could be argued that a weakness of this study is that we only used 2 radiologists, the radiologists were, in fact, independent of one another and worked in different countries. Furthermore, they were blind to each other’s results and had no involvement in developing the protocol. The fact that they achieved almost identical outcomes in these circumstances argues for the generalizability of the protocol.

Because of time constraints, we did not have the radiologists examine the same scan more than once. Thus, we were unable to measure the intraradiologist agreement for the protocol directly. Intraobserver reliability is usually measured before measuring that between observers. This is because a measurement that cannot be successfully repeated by an individual would be unlikely to have good reliability between different individuals. If we had seen low agreement between the 2 radiologists, a step in examining the cause would have been to do a test-retest study. Because we had such excellent agreement between our raters, this step seemed unnecessary. It is virtually impossible that such high interrater reliability could be achieved in the face of low intrarater reliability.

By using the protocol, the radiologists had almost perfect agreement on the selection of the cuts to be used as a basis of the Haller index. This finding has strong practical implications for use of the protocol to help reduce the
potential for variation when more than 1 clinician is calculating a Haller index. The protocol also provides a means for calculating an asymmetry index that, we expect to show, is useful in predicting outcome after surgical repair. In addition, use of the cross-sectional area is less likely to be impacted by the shape of the chest than any currently used index. The protocol provides detailed anatomical representation of the pectus deformity, which is important for determining surgical necessity. It is hoped that this more detailed picture of the deformity ultimately will be more useful in predicting physiological outcome than the single indices currently in use.

References


