Indirect determination of serum creatinine reference intervals in a Pakistani pediatric population using big data analytics

Sibtain Ahmed  
*Aga Khan University*, sibtain.ahmed@aku.edu

Jakob Zierk  
*University Hospital Erlangen, Erlangen, Germany*

Imran Siddiqui  
*Aga Khan University*, imran.siddiqui@aku.edu

Aysha Habib Khan  
*Aga Khan University*, aysha.habib@aku.edu

Follow this and additional works at: https://ecommons.aku.edu/pakistan_fhs_mc_pathol_microbiol

Part of the Pathological Conditions, Signs and Symptoms Commons, Pathology Commons, and the Pediatrics Commons

**Recommended Citation**

Available at: https://ecommons.aku.edu/pakistan_fhs_mc_pathol_microbiol/1342
Indirect determination of serum creatinine reference intervals in a Pakistani pediatric population using big data analytics

Sibtain Ahmed, Jakob Zierk, Imran Siddiqui, Aysha Habib Khan

BACKGROUND

The indirect methods of reference intervals (RI) establishment based on data mining are utilized to overcome the ethical, practical challenges and the cost associated with the conventional direct approach.

AIM

To generate RIs for serum creatinine in children and adolescents using an indirect statistical tool.

METHODS

Data mining of the laboratory information system was performed for serum creatinine analyzed from birth to 17 years for both genders. The timeline was set at six years from January 2013 to December 2018. Microsoft Excel 2010 and an indirect algorithm developed by the German Society of Clinical Chemistry and Laboratory Medicine’s Working Group on Guide Limits were used for the data analysis.

RESULTS

Data were extracted from 96104 samples and after excluding multiple samples for the same individual, we calculated RIs for 21920 males and 14846 females, with stratification into six discrete age groups.

CONCLUSION

Serum creatinine dynamics varied significantly across gender and age groups.

Key Words: Creatinine; Pediatric; Reference intervals; Indirect; Data mining; Pakistan
©The Author(s) 2021. Published by Baishideng Publishing Group Inc. All rights reserved.

Core Tip: Good laboratory practices advocate the necessity for generation of population specific reference intervals (RIs). The indirect methods of RIs establishment based on data mining are utilized to overcome the ethical, practical challenges and the cost associated with the conventional direct approach. The population specific RIs generated for pediatric serum creatinine levels in this study will assist in more accurate comprehension of the variations in creatinine and facilitate patient care.

INTRODUCTION

Reliable, accurate and population specific reference intervals (RIs) for laboratory analyses are pivotal for laboratory results interpretation and appropriate clinical decision-making. RIs for an analyte are based on the 2.5th and 97.5th centiles values from a set of pre-defined healthy individuals[1,2]. Furthermore, to improve the diagnostic efficiency of biomarkers, various partitioning criteria for RIs have been developed, particularly aimed to evaluate the influence of increasing age and gender dependence[3,4]. In the pediatric population, this portioning becomes more essential as physiological developments after birth and during adolescence result in fluctuations in the levels of many biomarkers, especially serum creatinine (CREA)[5].

The most commonly utilized and recommended ‘direct approach’ for RIs generation follows a more robust strategy, with a pre-selected population, that undergoes sample collection, processing and analysis in a controlled environment[6]. However, to utilize this approach in pediatrics is a challenging task, owing to ethical, financial and practical issues. Whereas, the indirect approach can be more effectively and conveniently utilized as an alternative route[6,7]. Analyte specific results from laboratory health records that comprise results obtained from healthy individuals as well as pathologic test results from clinical care areas are extracted in the indirect method and no additional blood samples are drawn, which is of utmost concern in children. This approach is swift and cost-effective especially for low middle-income countries (LMIC). Moreover, use of a minimum of 400 reference subjects for each partition aimed at obtaining statistically reliable RI calculations is further recommended, which can be conveniently accomplished with this approach[8].

In most clinical settings, evaluation of kidney function is carried out by requisition of biochemical analysis of serum CREA and 24 h CREA clearance as an indirect measure for the estimation of glomerular filtration rate (GFR)[9]. However, the growth mediated changes in CREA, especially in infancy and during puberty, due notably to its renal tubular secretion and the influence of muscle mass and dietary intake, makes the interpretation even more challenging[10].

The majority of laboratories in LMIC, are unable to establish their population specific RIs and seldom rely on published literature or adopt the ones cited by the manufacturers in kit information sheets[11]. Whereas, some laboratories also implement RIS calculated based on different analytical platforms and reagents than the ones in actual use. The inappropriate RIs adopted can lead to errors in report interpretation, ultimately leading to compromised patient safety, unnecessary further testing and costs, especially for LMIC. Our primary objective was to establish gender- and age-specific RIs for CREA specific to Pakistani children and adolescents using a validated indirect statistical approach[5,7,12].
MATERIALS AND METHODS

Study design and subjects
A team of investigators performed data mining of the laboratory information system at the Section of Clinical Chemistry, Aga Khan University. Ethical approval for the study was obtained from the Ethical review committee (ERC, #5348-Pat-ERC-18) of the university. All serum CREA measurements for both genders, including both in-house as well as ambulatory cases from birth to 17 years, were retrieved, regardless of the indication for test requisition. The timeline was set at six years from January 2013 to December 2018.

Biochemical analysis
The biochemical analysis was carried out on a Siemens ADVIA 1800 platform. The precision of the assay was 3.8% at 1.8 mg/dL (159 μmol/L) and 3.7% at 8.4 mg/dL (743 μmol/L), and the method was linear from 0-25 mg/dL (0-2210 μmol/L). As most of the laboratories in Pakistan are well versed with the conventional system of units, the levels of CREA are expressed in mg/dL. The laboratory is accredited by the College of American Pathologist and internal quality assurance is practiced in light of the Clinical & Laboratory Standards Institute standards.

Statistical analysis
The statistical analysis was performed using Microsoft Excel 2010 and the indirect algorithm proposed and pre-validated the German Society of Clinical Chemistry and Laboratory Medicine’s Working Group freely available online as a software pack-age [5,7,12]. The method is based on utilizing an input dataset of laboratory values containing both non-pathologic and pathologic samples, but only one sample per patient. A Power Normal distribution, defined as Gaussian distribution following Box-Cox transformation was performed to model the distribution of non-pathologic samples in the dataset. As per the default settings, the abnormal values are expected outside the distribution of normal CREA results, with an adjustment of the algorithm for the generation of the upper limits of the RI, by setting the Pathological value to “high”, compared to the physiological test results.

To calculate the respective 2.5th and 97.5th percentiles, the data were split into six age groups, for each gender, ranging from birth i.e., 0 d- < 2 years, 2- < 5 years, 5- < 9 years, 9- < 12 years, 12- < 15 years and 15- < 17 years, respectively, as defined previously by Tahmasebi et al[11] in the CALIPER cohort of healthy children and adolescents[11].

For the evaluation of calculated RIs, we performed a comparison of our results with Tahmasebi et al[11] that has established pediatric RIs for CREA on the Siemens ADVIA 1800[11]. Additionally, we also compared our findings with a local study by Molla et al [13] that has established direct RIs for CREA in an apparently healthy Pakistani population, for the combined 0-14 and 15 years onwards age groups, respectively, without partitioning into fine grained age clusters[13]. Lastly, the RIs currently in use by our laboratory for children and adolescents adopted from the Tietz textbook of clinical chemistry and molecular diagnostics were also evaluated[14].

RESULTS
From a total of 96104 samples analyzed for CREA over the study timeline, patients with multiple samples were further scrutinized and only the first sample analyzed was included in the final analysis. The lower and upper RIs were calculated based on 36766 CREA results obtained, including 21920 males and 14846 females as depicted in Tables 1 and 2. The complex age-related dynamics were more pronounced in the pre-pubertal group as represented by a significant proportion of samples in this age range.

Figures 1 and 2 illustrate the comparison of our results with RIs established using the direct method as reported by Tahmasebi et al[11], Molla et al[13] and the current RIs being used for reporting by our laboratory adopted from the Tietz textbook of clinical chemistry and molecular diagnostics.

DISCUSSION
Due to the lack of standardized data formats and experience in dealing with big data analytics, the majority of laboratories in LMIC as well as a few developed countries,
considerably lag behind in evaluating the transformative potential of the big data they have in store. The methodology employed was based on big data analytics and extraction of data from the laboratory information system of a tertiary care hospital’s laboratory that receives specimens from the entire country in order to ensure participation from all the ethnic groups existing in Pakistan.

Compared to the study by Molla et al[13] and RIs reported in the Tietz textbook of clinical chemistry and molecular diagnostics, a notable strength of this study is that it demonstrates a strong influence of age on CREA activity with the age-wise partitioning of RIs[12,13]. The differences noted, adds strength to the fact that it is imperative in clinical care to use age- and gender-specific RIs, for adequate comprehension of the dynamics of this widely used renal biomarker[5].

A literature review revealed that most of the reported RIs for CREA, have been established using healthy population-based approaches i.e. direct methods. While this approach is undoubtedly considered the gold standard, it has certain limitations.

### Table 1 Distribution of lower and upper reference intervals of creatinine in Pakistani male children

<table>
<thead>
<tr>
<th>Age (yr)</th>
<th>n</th>
<th>Our study, LRI (mg/dL)</th>
<th>Our study, URI (mg/dL)</th>
<th>Tahmasebi et al[11], LRI (mg/dL)</th>
<th>Tahmasebi et al[11], URI (mg/dL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 2</td>
<td>9658</td>
<td>0.15 (13 μmol/L)</td>
<td>0.39 (34 μmol/L)</td>
<td>0.17 (15 μmol/L)</td>
<td>0.52 (46 μmol/L)</td>
</tr>
<tr>
<td>2- &lt; 5</td>
<td>2964</td>
<td>0.15 (13 μmol/L)</td>
<td>0.80 (71 μmol/L)</td>
<td>0.18 (16 μmol/L)</td>
<td>0.49 (43 μmol/L)</td>
</tr>
<tr>
<td>5- &lt; 9</td>
<td>2833</td>
<td>0.16 (14 μmol/L)</td>
<td>0.69 (61 μmol/L)</td>
<td>0.31 (27 μmol/L)</td>
<td>0.61 (54 μmol/L)</td>
</tr>
<tr>
<td>9- &lt; 12</td>
<td>1796</td>
<td>0.27 (24 μmol/L)</td>
<td>0.92 (81 μmol/L)</td>
<td>0.41 (36 μmol/L)</td>
<td>0.68 (60 μmol/L)</td>
</tr>
<tr>
<td>12- &lt; 15</td>
<td>2291</td>
<td>0.29 (26 μmol/L)</td>
<td>1.06 (94 μmol/L)</td>
<td>0.43 (38 μmol/L)</td>
<td>0.85 (75 μmol/L)</td>
</tr>
<tr>
<td>15- &lt; 17</td>
<td>2378</td>
<td>0.40 (35 μmol/L)</td>
<td>1.26 (111 μmol/L)</td>
<td>0.61 (54 μmol/L)</td>
<td>1.02 (90 μmol/L)</td>
</tr>
</tbody>
</table>


### Table 2 Distribution of lower and upper reference intervals of creatinine in Pakistani female children

<table>
<thead>
<tr>
<th>Age (yr)</th>
<th>n</th>
<th>Our study, LRI (mg/dL)</th>
<th>Our study, URI (mg/dL)</th>
<th>Tahmasebi et al[11], LRI (mg/dL)</th>
<th>Tahmasebi et al[11], URI (mg/dL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 2</td>
<td>6323</td>
<td>0.12 (11 μmol/L)</td>
<td>0.73 (65 μmol/L)</td>
<td>0.17 (15 μmol/L)</td>
<td>0.52 (46 μmol/L)</td>
</tr>
<tr>
<td>2- &lt; 5</td>
<td>2012</td>
<td>0.15 (13 μmol/L)</td>
<td>0.74 (65 μmol/L)</td>
<td>0.18 (16 μmol/L)</td>
<td>0.49 (43 μmol/L)</td>
</tr>
<tr>
<td>5- &lt; 9</td>
<td>1997</td>
<td>0.16 (14 μmol/L)</td>
<td>0.68 (60 μmol/L)</td>
<td>0.31 (27 μmol/L)</td>
<td>0.61 (54 μmol/L)</td>
</tr>
<tr>
<td>9- &lt; 12</td>
<td>1204</td>
<td>0.26 (23 μmol/L)</td>
<td>0.78 (69 μmol/L)</td>
<td>0.36 (32 μmol/L)</td>
<td>0.63 (56 μmol/L)</td>
</tr>
<tr>
<td>12- &lt; 15</td>
<td>1573</td>
<td>0.24 (21 μmol/L)</td>
<td>0.84 (74 μmol/L)</td>
<td>0.40 (35 μmol/L)</td>
<td>0.72 (64 μmol/L)</td>
</tr>
<tr>
<td>15- &lt; 17</td>
<td>1737</td>
<td>0.34 (35 μmol/L)</td>
<td>0.93 (82 μmol/L)</td>
<td>0.50 (44 μmol/L)</td>
<td>0.77 (68 μmol/L)</td>
</tr>
</tbody>
</table>


### Figure 1 Comparison of serum creatinine reference intervals in males. LRI: Lower reference limit; URI: Upper reference limit.
including those specifically pertaining to expenses for conducting these large-scale prospective studies especially for a LMIC. Additionally, attainment of a minimally acceptable sample size for the different age groups in pediatrics is also a concern. The indirect method not only made it possible to statistically analyze big data \((n = 36766)\), acquired as part of routine care, which further minimized the ethical and practical concerns. However, this approach, requires significant refinement of the specimen selection alongside validated and robust statistical analysis. In this context, we utilized an established algorithm that had already been extensively evaluated and validated by large scale multicenter studies\([4,15]\). Notably, a literature review revealed that RIs in children established using direct methods do not correctly account for the extensive changes with age as most of them lack age-based partitioning. Moreover, in instances of non-normal distribution, the direct method often generates unacceptably broad confidence intervals (CIs) limiting their widespread adoption\([16]\).

Next to, the RIs reported in the CALIPER cohort, our proposed RIs for CREA seem to differ. In particular, our lower reference limits (LRIs) are considerably lower than the CALIPER cohort, indicating that Pakistanis tend to have a different genetic structure with significantly lower lean tissue mass and a lower GFR compared to the CALIPER cohort. The LRIs and upper reference limits from the CALIPER cohort and the study by Molla \textit{et al}\([13]\) remain continuous up to five years of age, on the contrary, this study demonstrates pronounced age-related fluctuations in this age group for both genders. The maximum values were attained at 12 years in all the studies evaluated, trailed by an incline, having a probable association with the increase in muscle mass with age and attainment of puberty. On gender stratification, our study demonstrated that the peak levels of CREA attained in males \(1.26 \text{ mg/dL} (111 \mu\text{mol/L})\) significantly differed from females \(0.93 \text{ mg/dL} (82 \mu\text{mol/L})\). The need for fine grained age- and gender-based RIs for CREA is also supported by another study by Pottel \textit{et al}\([17]\) that has established age- and gender-specific CREA RIs from hospital laboratory data based on different statistical methods, and has shown pronounced age-based fluctuation in CREA for both genders\([17]\). This phenomenon is in accordance with the dependency of CREA on physical structure, muscle mass, physical activity and protein uptake which differs significantly between the two genders\([18,19]\). Furthermore, as the utilized method does not allow creation of CIs, equivalence limits were derived according to previously established and validated equations and significant differences between our study RIs and Tahmasebi \textit{et al}\([11]\) were noted as depicted in Tables 1 and 2. It is evident the direct and indirect methods can more often generate overlapping but not identical values\([20]\).

Considering the scarcity of literature on fine grained age group-based pediatric RIs for CREA in Pakistan, one of the highly densely populated countries reportedly with a high burden of kidney disease, the data mining approach can serve as the missing link\([21,22]\). Furthermore, the deployment of indirect approaches using “big data” solutions are barely utilized in LMIC and this study highlights the utility of this approach at no additional cost. Several LMIC lack a medical insurance system with universal coverage; thus, in most instances, the expenditure has to be self-born by the subjects\([23]\). Adequate interpretation based on population specific RIs can prevent unnecessary further investigations and medical interventions\([24]\). This study is in line with good laboratory practices that advocate the need for RIs establishment alongside the attainment of the quality improvement of the post analytical phase, aimed at appro-
appropriate report interpretation.

In addition to the merits of this real-world big-data approach in laboratory medicine, there is a notable limitation of this indirect algorithm, that any potential differences cannot be analyzed between the groups formulated; hence, individual results have to been complemented with clinical judgement and correlation. Moreover, the CIs with the established RIs were not calculated, as the used algorithm does not contain a provision for CI generation.

**CONCLUSION**

Good laboratory practices advocate the necessity for generation of population specific RIs, which is widely lacking, particularly in LMIC owing to the various challenges of the conventional direct method. This study has highlighted and further substantiated the utility of an alternative validated indirect algorithm by data mining in a clinical laboratory in Pakistan. This approach can be easily adopted by laboratories in resource constrained regions and the RIs generated will provide more accurate comprehension of laboratory reports in order to facilitate clinical care.

**ARTICLE HIGHLIGHTS**

**Research background**
Population specific reference intervals (RIs) are pivotal for laboratory results interpretation.

**Research motivation**
The indirect methods of RIs establishment based on big data analytics overcome the challenges and the cost associated with the conventional direct approach.

**Research objectives**
To establish RIs for serum creatinine (CREA) levels in Pakistani children using an indirect data mining approach.

**Research methods**
RIs were calculated using a previously validated indirect algorithm developed by the German Society of Clinical Chemistry and Laboratory Medicine’s Working Group on Guide Limits.

**Research results**
The lower and upper RIs were calculated based on 36766 CREA results obtained from 21920 males and 14846 females.

**Research conclusions**
These RIs generated for serum CREA demonstrate the complex age- and gender-related dynamics occurring with physiological development.

**Research perspectives**
This indirect approach can be easily adopted by laboratories in resource constrained regions and the RIs generated will provide more accurate comprehension of laboratory reports in order to facilitate clinical care.

**REFERENCES**


