Data Mining in Establishing the Indirect Reference Intervals of Biochemical and Haematological Assays in the Paediatric Population: A Review

Dian N. Nasuruddin[1](✉), Ely Salwana[1], Mahidur R. Sarker[1], Adli Ali[2], and Tze Ping Loh[3]

[1] Institute of Visual Informatics, Universiti Kebangsaan Malaysia, Bangi, Malaysia
p112770@siswa.ukm.edu.my
[2] Paediatrics Department, Pusat Perubatan Universiti Kebangsaan Malaysia, Kuala Lumpur, Malaysia
[3] Department of Laboratory Medicine, National University Hospital, Lower Kent Ridge Road, Singapore

Abstract. Reference intervals (RIs) are fundamental values accompanying medical laboratory results that allow interpretation by medical practitioners, thus influencing patient management. Traditionally, RIs are established by recruiting 120 healthy reference individuals and applying statistical analysis to the results. This method is challenging due to the technical and ethical issues involved. Therefore, many laboratories either adapt RIs provided by the manufacturers of their analytical platforms or the results of RI studies done in other countries. The advent of data mining technology has allowed an alternative method, the indirect RIs (IRIs) approach, which applies appropriate statistical techniques to patient data stored in the laboratory electronic medical records to establish the IRIs. This review briefly highlights the historical aspect of IRI determination, provides a general outline of the steps involved and reviews publications that have used data mining to establish the paediatric IRI over the past ten years.

Keywords: Data mining · Reference intervals · Indirect approach · Continuous centile curves · Paediatrics

1 Introduction

The application of data mining in the medical domain, including the diagnostic medical laboratory, has recently seen an increasing trend. Data mining can be broadly defined as 'a set of mechanisms and techniques, realised in software, to extract hidden information from data' [1]. It is a subprocess of knowledge discovery in data (KDD) which is the 'non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data' [2]. The technical approaches of applying data mining in the medical world include data clustering and classification, making predictions, finding

H. Badioze Zaman et al. (Eds.): IVIC 2023, LNCS 14322, pp. 493–507, 2024.
https://doi.org/10.1007/978-981-99-7339-2_41

frequent patterns, analysing changes, and detecting anomalies. Clinical laboratory test results are paramount to evidence-based medicine, with nearly 80% of medical decisions made on the information provided by laboratory reports [3]. Without an accompanying set of reference intervals (RIs), a test result on its own is of little value [4]. A reference interval (RI), as defined by Ceriotti [5], "is an interval that, when applied to the population serviced by the laboratory, correctly includes most of the subjects with characteristics similar to the reference group and excludes the others."

The RI serves as a health-associated benchmark with which to compare an individual test result and is vital in the implementation of mobile health monitoring system (mHealth) as we usher in the 4.0 industrial revolution. This system would enable clinicians and empower patients by illustrating the trace of critical physiological parameters, generating early warnings/alerts, and indicating the need for any significant changes to the results, consultation, medication, and treatments [6]. However, establishing accurate and reliable RI is considerably complex [7]. The paediatric RIs (PRIs) should reflect the dynamic biological and biochemical changes throughout the developmental growth to ensure correct diagnosis and treatment [8].

While the concept of RIs and their values appears simple, defining paediatric reference intervals (PRIs) using the direct method involving 120 presumed healthy reference individuals per partition is daunting and taxing. The cost of conducting a direct reference interval study based on the activity-based-costing (ABC) method described is also high [9]. Due to the obstacles accompanying the establishment of PRIs using the direct method, an alternative which is the indirect method, has started to garner a lot of attention. The indirect reference interval (IRI) method involves data mining of routine paediatric laboratory results collected for other purposes, including routine clinical care and screening from the laboratory information system (LIS). By using appropriate statistical techniques, PRIs are subsequently established [10]. The primary clinical data mining method used in IRI is descriptive cluster analysis, which is finding similar groups of objects to form clusters. It is an unsupervised machine-learning-based algorithm that acts on unlabeled data. A group of data points would comprise a cluster in which all the objects would belong to the same group, i.e., the partitioning of similar data points. The clustering methods commonly used include partitioning and density-based methods.

The IRI is based on identifying a distribution amid the data and does not require assessment of all individual results in the dataset as belonging to the reference population [10]. The IRI method assumes that the examined dataset consists of a mixture of parametrically distributed samples from healthy individuals and pathological samples not described by that distribution. In a sufficiently large dataset with a dominant fraction of physiological test results for the examined analyte, the distribution of non-pathological values can be estimated using advanced statistical methods and the pathological test results are assumed to have no substantial impact on the RIs [11].

This paper aims to highlight the historical aspect of IRI determination and the assessment of publications that have used data mining to establish the IRI in the pediatric population over the past ten years. This paper is arranged into five sections. Section two briefly explains the data mining in indirect PRIs establishment. This is followed by a summary of selected articles that utilised data mining to establish PRIs in Sect. 3, a discussion of the results in Sect. 4 and ending with the conclusion in Sect. 5.

2 Data Mining in IRI Establishment

This section describes the historical aspect of data mining in IRIs and the general steps involved in its establishment.

2.1 The Indirect Methods for RIs Establishment

The foundation of establishing IRIs using patients' results stored in the LIS was laid as early as the 1960s by Robert G Hoffman, who proposed the application of statistics in medicine in the Journal of the American Medical Association [12] and was documented initially by John Glick in 1972. However, it was not until the personal computer arrived in the 1980s that enough computing power was available to apply it generally [13]. Subsequently, C. G. Bhattacharya explored a graphical method to identify Gaussian distribution components in 1967 [14]. This has paved the way for other scholars to apply the method in their research.

T Kouri and his team developed RIs for haematological blood indices partitioned for gender by combining data mined from the LIS and diagnostic data. They surmised that data mined from hospitalised patients based on diagnostic information may apply to other analytes. Horn and Pesce 1998 developed a robust approach for establishing RIs for small datasets. [15]. Later in 2019, a modified version was presented by Horn et al. to accommodate larger distributions of reference intervals [16]. The REALAB project by Grossi et al. in 2005 established RIs for 23 basic tests using approximately 15 million records using a multivariate algorithm. [17]. A novel approach of using a kernel-smoothed density function based on a bimodal method to estimate the distribution of the combined data for both non-diseased and diseased populations was developed by Arzideh et al. in 2007. This is a more advanced procedure to determine RIs from data mined in laboratory databases without considering any diseased population distribution. [18]. The Clinical Laboratory Standard Institute (CLSI), in its 2010 guidelines issue, has guided the establishment of RIs for quantitative clinical laboratory tests. The indirect method was briefly mentioned as an alternative, not a primary one, to replace direct RIs. With the explosion of big data technology, the challenges in recruiting reference individuals and the exorbitant cost involved in developing RIs using the direct method, especially in the paediatric population, many researchers from around the globe have taken an interest in exploring and conducting studies focusing on developing PRIs from data mining of patient data from diagnostic laboratories using the indirect method. This has led to the improvement of the methodology and statistical techniques in leaps and bounds [18–22].

2.2 The Indirect Methods for Paediatric RIs (PRIs) Establishment

A long-standing gap exists in the PRIs, especially in the neonates and young infant subgroups. This is because of the difficulty and ethical issues in obtaining blood from the healthy paediatric population. With the growth of technology and the availability of large laboratory databases, the indirect reference interval method is seen to have the full potential to fill in this gap. The challenge is determining the physiological samples amidst the pathological samples in the mixed laboratory dataset using either the metadata-driven or primary statistical strategy. The availability of many data set points has also contributed to the development of continuous percentile charts or dynamic reference intervals of biochemical and haematological analytes, which better represent the dynamic physiological development in the paediatric population, especially during the neonatal/infantile period and throughout puberty [23, 24]. The indirect method has mainly been used to establish paediatrics IRIs (PIRIs) for biochemical analytes such as calcium and bone markers, alkaline phosphatase, creatinine, lipids, arterial blood gases, creatinine and trace minerals [22, 25–33]. Apart from that, the indirect method has also been successfully used to establish haematological reference intervals for full blood count indices and coagulation profiles in many countries [21, 34–39].

2.3 The Steps Involved in IRI Establishment

The general steps involved in the IRI establishment, whether metadata driven or statistically-driven, include data collection, cleaning, data analysis and result verification. Data analysis comprises three main processes: partitioning of the input dataset according to desired groups, statistical analysis, which includes outlier removal, calculation of cumulative frequency (cdf) of each result, calculation of the inverse cdf of a standard Gaussian distribution and graphing the inverse cdf versus each of the measured analyte value and performing piece-wise linear regression in R software to identify the linear portion of the distribution. This is followed by graphing the linear part of the distribution and using linear regression to determine the equation that represents the linear portion of the distribution. Next, by using the linear equation, the 2.5th and 97.5th centiles may be extrapolated and taken as the lower and upper reference intervals [40]. In metadata-driven studies, additional steps are taken in the data cleaning process to remove results associated with abnormality of other analytes from any patients with known diseases, or the opinion of subject matter experts.

In calculating continuous reference intervals, additional steps would need to be taken. This involves dividing the datasets into overlapping timeframes, excluding pathological values using statistical methods, and calculating the 2.5th, 50th and 97.5th percentiles of the remaining values of each parameter using statistical and graphical software such as R and R Studio. Special consideration in the IRI is the verification of the various indirect approaches. To verify the newly established IRIs, many researchers may directly compare the results with previously published articles in the literature to assess the agreement or may perform an in-house verification using the standard verification procedure described by CLSI EP28-A3c, which emphasises that three approaches that can be used to verify RIs, i.e. subjective assessment, using a small number ($n = 20$) of reference individuals or using a large number or reference individual ($n = 60$ but fewer than 120). In the second and third approaches, if no more than 2 of the 20 samples (i.e., 10% of the test results) fall outside the RI, at least provisionally, it may be received for use. However, if 3 or 4 of the 20 samples fall outside the RI, a second set of 20 reference specimens should be obtained, and if again three or more of the new specimens (i.e., $\geq$ 10% of the test results) OR 5 or more of the original 20 falls outside the RI, the user should re-examine the analytical procedures used and consider possible differences in the biological characteristics of the two populations sampled.

At the time of writing this article, there are many published algorithms for derivation of IRI. Among them include the Hoffman and the modified Hoffman methods, the Bhattacharya method, the Arzideh method, and the Wosniok method [41]. Simulation studies are highly recommended in comparing the various indirect methods' diagnostic efficiency and allow appropriate statistical confidence analysis [42].

3 Summary of Published Studies on RI Establishment Using Indirect Method in the Paediatric Population

Tables 1, 2 and 3 summarise the studies that employed the indirect method to establish paediatric reference intervals. Three databases (Scopus, EBSCO Medline and WOS) were searched using the terms' data mining', 'data analytics', 'big data', 'calculating', 'constructing', 'developing', 'establishing', 'reference interval', 'normal range', 'reference limit', 'reference curves', 'paediatrics', 'child', 'adolescent', 'newborn', and 'neonate' from 2012 through July 2022.

Table 1 presents a detailed summary of published papers reporting the establishment of PRIs of biochemical assays by indirect methods. This study will compare selected studies based on a few criteria, including the year, the country in which the study was conducted, the analytes included in the study, the methods used, discrete vs continuous PRIs and the type of partitioning established. Ten studies were included from 2012 through 2022.

In 2012, Eduardo et al. from Argentina established discrete age-specific thyroid hormones IRIs using laboratory results over a period of 5 years involving 7581 children [43, 44]. This study was meta-data driven as rigorous exclusion criteria were applied to the data prior to the final analysis. This study established higher TSH and T4 values than a previous direct RI study done in German [44], highlighting the importance of population-specific RIs. In the same year, a group of researchers from Israel established their discrete IRI partitioned by age for TSH and free T3 using results from over 11,000 children and adolescent and found that the then RI used were too low and suggested the transference of their results to other laboratories [45]. There was no partitioning based on gender done for both studies. Another study in the UK published in 2013 successfully established age and gender-specific IRIs for serum prolactin to aid in diagnosing neurometabolic conditions affecting dopamine metabolism [46]. This study extracted over ten years of data from 2369 hospital patients. The established IRI was comparable with previously published IRIs [47] and has filled the knowledge gap by providing the prolactin RI for infants under one year. In the same year, a group of researchers in America established the discrete age-specific IRIs for calcium using 4629 datasets. This meta-data-driven study found that the calcium IRIs were broader than the currently used and suggested that the differences may reflect seasonal or ethnic heterogeneity [48].

The Canadian group in 2014 published a paper studying the validity of establishing PRIs based on hospital patient data by comparing the age and gender-specific discrete PRIs results of 13 biochemical analytes established using the indirect method (modified Hoffmann) to results obtained in the CALIPER study [40]. This statistically-driven study analysed over 200,000 data points per analyte and found that the indirect PRIs established were generally wider than the CALIPER study. Another single-centre, metadata-driven study in Turkey published in 2015 analysed 1709 data points and developed gestational age-specific TSH and free T4 continuous IRIs. They found that free T4 correlated with gestational whilst TSH remained unchanged irrespective of gestational age [49]. In the same year, a team of researchers from Denmark published the results of their multicentre, statistically-driven study that analysed the creatinine results of over 11,000 data sets. The continuous age and gender-specific IRIs showed that age dependency was seen in both boys and girls from birth to adulthood [50].

A large multicentre, statistically-driven study in the Netherlands published in 2019 [51] analysed 7,574,327 results of children visiting their general practitioners and established discrete age and gender-specific IRIs of 18 biochemical analytes with the aim to adapt them as standardised national RIs. They found that there were significant age effects for liver enzymes and creatinine. One single-centre, statistically-driven study done in Pakistan was published in 2021. The group analysed 96104 data points and established discrete IRI for creatinine and found that the serum creatinine dynamics differ across gender and age groups. Compared to CALIPER, their creatinine IRIs were lower. This is thought to be due to the different genetic structures and, again, highlights the importance of developing population-specific RIs. Another large statistically-driven multicentre study in Germany established high-resolution age and gender-specific continuous IRIs for 15 biochemical analytes using an analysis of 217, 883 - 982,548 samples per analyte which showed high concordance to the continuous RIs of other large direct studies (CALIPER and HAPPi Kids) [52].

Table 2 presents a detailed summary of published papers reporting the establishment of PRIs of haematological and coagulation assays by indirect methods. Six studies are included. The first study was done in Romania and published in 2013. This group of researchers conducted a single-centre, meta-data-driven study of 845 patient data sets to establish discrete IRIs for erythrocyte parameters specific for one-day-old neonates [34]. They found that the results were comparable to previously published direct RIs [53]. The same team later in the following year published an article on the discrete IRIs for platelet parameters in the first day of life, neonates, using 1124 patient datasets and partitioning the results according to gender [54]. The obtained values for some parameters agreed with the literature, while some differed [55]. This supports the need for establishing population-specific haematology reference intervals.

Zierk et al. in 2013 published the results of a statistically-driven German single-centre study of age-specific continuous IRIs using analysis of 56,253 – 60,394 data points for various haematology indices [35]. In this study, the results were comparable to the previously published KiGGS study and managed to capture biological events. Then in 2018, Weidhofer et al. from Australia established continuous age and gender-specific IRIs for coagulation parameters [36]. This study extracted data from two centres and analysed 19,684–55,101 data sets. The resulting IRIs highlighted the coagulation parameters' age-dependent dynamics, and some of the parameters showed concordance with previous literature [56]. In 2019, Zierk and his team of researchers published the results of a large German metadata-driven multicentre study that analysed 9,576,910 samples from 358,292 patients that established continuous percentile charts of various haematology parameters partitioned according to age and gender [37]. They observed complex age and sex-related dynamics in haematology analytes during all periods of

Table 1. Published papers reporting the establishment of PRIs of biochemical assays by indirect methods

No	Author(s), year, country	Title	Discrete/Continuous, Partitioning
1	E. A. Chaler et al. (2012), Argentina [43]	Age-specific thyroid hormone and thyrotropin reference intervals for a pediatric and adolescent population	Discrete, Age
2	Strich, D. et al. (2012), Israel [45]	Current normal values for TSH and FT3 in children are too low: evidence from over 11,000 samples	Discrete, Age
3	Aitkenhead, H. et al. (2013), United Kingdom [46]	Establishment of paediatric age-related reference intervals for serum prolactin to aid in the diagnosis of neurometabolic conditions affecting dopamine metabolism	Discrete, age and gender
4	Roizen, J. et al. (2013), USA [48]	Determination of reference intervals for serum total calcium in the vitamin D-Replete pediatric population	Discrete, Age
5	Shaw, J. et al. (2014), Canada [40]	Validity of establishing pediatric reference intervals based on hospital patient data: A comparison of the modified Hoffmann approach to CALIPER reference intervals obtained in healthy children	Discrete, age and gender
6	Imamoglu, E. et al. (2015), Turkey [49]	Nomogram-based evaluation of thyroid function in appropriate-for-gestational-age neonates in intensive care unit	Continuous, Gestational age specific
7	Søeby, K. et al. (2015), Denmark [50]	Mining of hospital laboratory information systems: a model study defining age- and gender-specific reference intervals and trajectories for plasma creatinine in a pediatric population	Continuous, age and sex

(*continued*)

Table 1. (*continued*)

No	Author(s), year, country	Title	Discrete/Continuous, Partitioning
8	Den Elzen, W. P. J et al. (2019), The Netherlands [51]	NUMBER: Standardised reference intervals in the Netherlands using a 'big data' approach	Discrete, age and sex
9	Ahmed, S. et al. (2021), Pakistan [30]	Indirect determination of serum creatinine reference intervals in a Pakistani pediatric population using big data analytics	Discrete, Age
10	Zierk, J. et al. (2021), Germany [52]	High-resolution pediatric reference intervals for 15 biochemical analytes described using fractional polynomials	Continuous, age and gender

childhood and adolescence. Compared to their previous work in 2013, the current IRIs was narrower and showed high concordance with the KiGGS study. Another group of researchers from Germany published another article in 2022 [21]. This metadata-driven study was done in Berlin and Brandenburg to establish discrete IRIs for various haematology parameters. A total of 27,554 patient datasets were analysed, and age, as well as sex-specific IRIs, were established. The IRIs from this study showed differences from previously published articles which might be explained by the different population distribution due to high foreign influx [57, 58]. This further reiterates the need for the establishment of population-specific reference intervals.

Table 3 summarises three published papers reporting the establishment of PRIs of various biochemical, haematological, coagulation and other multi-discipline assays by indirect methods. Six studies are included. The first study was done in Germany by Zierk and his team of researchers [22]. This single-centre statistically driven study established the age and sex-dependent continuous reference intervals for 13 biochemical analytes and haematological parameters. In their research, electrolytes and total protein showed age-specific changes but not sex-specific. One of the analytes studied, alkaline phosphatase, showed complex dynamic patterns, and most of the analytes' IRIs were comparable to CALIPER and KiGGS studies.

Table 2. Published papers reporting the establishment of PRIs of haematological assays by indirect methods

No	**Author(s), year, country**	**Title**	**Discrete/Continuous**, Partitioning
1	Grecu, D. S., et al. (2013), Romania [34]	Quality in post-analytical phase: indirect reference intervals for erythrocyte parameters of neonates	Discrete, Age-specific for one-day-old neonates
2	Zierk, J. et al. (2013), Germany [35]	Indirect determination of pediatric blood count reference intervals	Continuous, Age
3	Grecu, D. S., et al. (2014), Romania [54]	Quality assurance in the laboratory testing process: indirect estimation of the reference intervals for platelet parameters in neonates	Discrete, Age (First day of life) and sex
4	Weidhofer, C. et al. (2018), Austria [36]	Dynamic reference intervals for coagulation parameters from infancy to adolescence	Continuous, age and gender
5	Zierk, J. et al. (2019), Germany [37]	Next-generation reference intervals for pediatric hematology	Continuous – percentile charts, age and gender
6	Mrosewski, I. et al. (2022), Germany [21]	Indirectly determined hematology reference intervals for pediatric patients in Berlin and Brandenbur	Discrete, age and sex-specific

A team from Korea presented the results of their large multicentre study in 2021. This metadata-driven study established the discrete age and gender-specific IRIs for haematology, biochemical and coagulation parameters. The PRIs determined from this study differed from existing results and PRIs from other ethnicities. Subsequently, a team of researchers from America also published their age and gender-specific discrete reference intervals for 266 individual analytes across multiple clinical disciplines [59]. Patient results from 13 laboratories amounting to a total of 71,594,330 total patients test results were analysed in this statistically-driven study, and the team has successfully established IRIs with very powerful sample sizes for each age bracket.

Table 3. Published papers reporting the establishment of PRIs of biochemical, haematological and coagulation assays by indirect methods

No	Author(s), year, country	Title	Discrete/Continuous, Partitioning
1	Zierk, J., et al. (2015), Germany [22]	Age- and sex-specific dynamics in 22 hematologic and biochemical analytes from birth to adolescence	Continuous, age and sex-dependent change during development
2	Sung, J. Y. et al. (2021), Korea [60]	Establishment of Pediatric Reference Intervals for Routine Laboratory Tests in Korean Population: A Retrospective Multicenter Analysis	Discrete, age and sex
3	Fleming, J. et al. (2022), USA [59]	Development of nationwide reference intervals using an indirect method and harmonised assays	Discrete, age and gender (neonatal, paediatric, adults, geriatric)

4 Summary of Published Studies on RI Establishment Using Indirect Method in the Paediatric Population

This narrative review provides a historical review of data mining in the paediatric IRI determination and an assessment of the published articles within the past years that have utilised data mining in establishing the paediatric indirect reference intervals over the past ten years. There are many advantages of using the indirect method compared to the direct method. Indirect methods harness the power of big data that increases statistical power, are representative of the true population and allow easy application of complex statistical analysis to be applied to thousands and even millions of deidentified data points pulled from the laboratory database of a single or many centres for fast outlier removal, transformation and partitioning to establish robust IRIs. Applying further statistical analysis would allow for the creation of continuous or dynamic percentile charts that better represent the fluid physiological changes seen in children. The indirect method also provides analysis of retrospective data of difficult-to-obtain samples such as body fluids, CSF, and amniotic fluid, as the steps involved are identical to the analysis of data for serum, plasma, or whole blood samples.

On the contrary, the direct method is tedious as it involves recruiting healthy reference individuals, which is hard to come by, especially in healthy paediatric populations. The limited sample size reduces the statistical power, and application to a larger population is debatable. The typically small number of results hinders the ability to partition the data; hence, only discrete RIs could be established for certain arbitrarily set age brackets. Samples from reference samples would need to be collected, processed, and stored for batch analysis which may take longer. This cycle may also introduce bias in the result. Ethical issues involved are among the more challenging hurdles, as researchers would need to obtain informed consent from the parents of the paediatric reference individuals to allow the collection of data and venepuncture to be conducted. It is also more expensive to conduct the direct method as it involves the cost of reference individuals' reimbursement, labour of testing and the cost of reagents and consumables.

Recently, there has been a significant increase in the number of publications on the indirect method, especially over the past five years. This signifies an interest in IRI establishment as an alternative to the laborious direct method. The boost of interest, especially of the laboratorians, to embark and report results of indirect reference interval studies is most likely contributed by the advantages of the indirect method discussed previously, coupled with the advanced database available in the laboratory, readily available volume of patient data stored for analysis plus easy access to statistical analysis tools developed by the previous group of researchers. Initially, it was noted that many of the earlier publications reviewed did not include a thorough description of the data mining and the statistical methods used in the IRI establishment. However, subsequent publications have included detailed step-by-step descriptions of that IRI establishment entails.

Jones et al. [10] have proposed a checklist of the minimum requirements for publication of IRI studies which include details of study design, a description of the population and the data source, a description of available records of preanalytical and analytical processes, the data set selection and filtering criteria, the description of the data set inclusive of number of samples, median, kurtosis and initial analysis of partitioning. The description of the statistical process inclusive of outlier detection, method and transformation, results of statistical analysis, comparison with other statistically reliable peer-reviewed published studies and final recommendations and discussion of the study would also need to be included. This would allow future researchers to understand the overall steps involved and critique the study to find any weaknesses, strengths, and opportunities for improvements before conducting their own population-specific indirect reference studies.

4.1 Comparison of Indirect PRIs Between Countries

Most of the reviewed articles were done in the European population. Only two studies were done in the Asian population (Pakistan and Korea). Data mining and indirect sampling have allowed multiple laboratories in a country or a region to conduct IRI studies using the same methodology and analytical platforms to establish common reference intervals. However, caution needs to be exercised. This is exemplified by the results of two studies in different regions in Germany that have shown variability in their IRI results. The difference might be due to the difference in the population, as one region is known to have a high foreign influx [21, 37]. Many other studies done in Europe have reported good agreement with previously reported RIs established by both the direct and indirect methods [35, 37]. However, variability is still seen especially wider values of certain analytes [40, 48]. The two studies done in the Asian continent [30, 60] have also come up with different IRIs than the ones currently used in their population and previously published PRIs from other countries. This serves as a reminder to laboratories from other countries, especially those with diverse multi-ethnic heterogeneous populations, to be cautious in the transference of IRI results from different countries and further highlights the necessity of establishing own population-specific reference intervals preferably partitioned according to age, gender, and ethnicity.

4.2 Discrete vs. Continuous IRIs

The centile charts are familiar among most health care providers and parents as they are used to assess their children's developmental growth. The application of centile charts in biochemical and haematological paediatrics RI has led to the transformation of discrete RIs to dynamic continuous percentile charts [61, 62]. Percentile RI charts enable the removal of the arbitrarily set age group partitions that may confuse the interpretation of results, especially in children between age group brackets [63]. The intuitive percentile charts allow the physiological patterns and dynamics of paediatrics analytes to be visually represented. Seven articles described in this review have developed continuous reference intervals for various biochemical, haematological and coagulation assays [22, 35–37, 49, 50, 52]. The majority were multicentre studies and were statistically driven. Most of the studies applied the Arzideh method [64, 65, 22, 35–37, 52] and the kosmic software [52] in the calculation of continuous RIs. Even though there is a move towards developing continuous percentile charts, one major hurdle remains. Currently, many laboratories' information systems are unable to incorporate advanced mathematical functions or graphical representations of patient results [3]. Hopefully, this obstacle will soon be overcome, and continuous reference percentiles can be integrated fully into clinical practice.

5 Conclusion

This paper observed that data mining techniques have been employed successfully in establishing PIRIs. Caution must be exercised during data cleansing as this process must be done thoroughly to ensure the voracity of the established PRIs. There is still a paucity of data regarding the PRIs based on different ethnicities. Many of the published PRIs were based on the Caucasian population and might not be suitable for the transference of PRIs to other medical diagnostic laboratories elsewhere. Therefore, many authors have highlighted the importance of establishing the age, sex and ethnicity-specific to the population. Many researchers are moving towards the establishment of dynamic continuous PRIs using a few recently published algorithms and programs that help to understand the physiological dynamic changes in paediatric biochemistry and complement age-specific RIs in the tracking, interpretation and application of the results in clinical patient management.

References

1. Coenen, F.: Data mining: past, present and future. Knowl. Eng. Rev. **26**(1), 25–29 (2011)

2. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery: an overview. In: Advances in Knowledge Discovery and Data Mining, Editor. AAAI Press/The MIT Press, Menlo Park (1996)
3. Hoq, M., et al.: Paediatric reference intervals: current status, gaps, challenges and future considerations. Clin. Biochem. Rev. **41**(2), 43–52 (2020)
4. Katayev, A., Balciza, C., Seccombe, D.W.: Establishing reference intervals for clinical laboratory test results: is there a better way? Am. J. Clin. Pathol. **133**(2), 180–186 (2010)
5. Ceriotti, F.: Establishing pediatric reference intervals: a challenging task. Clin. Chem. **58**(5), 808–810 (2012)
6. Mat Nayan, N.: Model for monitoring chronic disease in MHealth applications: a case of Asian country. IRAJ (2020)
7. Tahmasebi, H., et al.: Pediatric reference intervals for biochemical markers: gaps and challenges, recent national initiatives and future perspectives. EJIFCC **28**(1), 43–63 (2017)
8. Lyle, A.N., et al.: Current state of pediatric reference intervals and the importance of correctly describing the biochemistry of child development: a review. JAMA Pediatr. **176**(7), 699–714 (2022)
9. Ibrahim, R., et al.: Estimation of cost of diagnostic laboratory services using activity based costing (ABC) for implementation of Malaysia diagnosis related group (MY-DRG®) in a teaching hospital. Malays. J. Publ. Health Med. **17**, 1–8 (2017)
10. Jones, G.R.D., et al.: Indirect methods for reference interval determination – review and recommendations. Clin. Chem. Lab. Med. (CCLM) **57**(1), 20–29 (2019)
11. Zierk, J., et al.: Indirect determination of hematology reference intervals in adult patients on Beckman Coulter UniCell DxH 800 and Abbott CELL-DYN Sapphire devices. Clin. Chem. Lab. Med. **57**(5), 730–739 (2019)
12. Yang, D., Su, Z., Zhao, M.: Big data and reference intervals. Clin. Chim. Acta **527**, 23–32 (2022)
13. Hoffmann, G., Lichtinghagen, R., Wosniok, W.: Simple estimation of reference intervals from routine laboratory data. Lab. Medizin **39**(6) (2016)
14. Bhattacharya, C.G.: A simple method of resolution of a distribution into gaussian components. Biometrics **23**(1), 115–135 (1967)
15. Horn, P.S., Pesce, A.J., Copeland, B.E.: A robust approach to reference interval estimation and evaluation. Clin. Chem. **44**(3), 622–631 (1998)
16. Beasley, C.M., Jr., et al.: Adaptation of the robust method to large distributions of reference values: program modifications and comparison of alternative computational methods. J. Biopharm. Stat. **29**(3), 516–528 (2019)
17. Grossi, E., et al.: The REALAB project: a new method for the formulation of reference intervals based on current data. Clin. Chem. **51**(7), 1232–1240 (2005)
18. Arzideh, F., et al.: A plea for intra-laboratory reference limits. part 2. a bimodal retrospective concept for determining reference limits from intra-laboratory databases demonstrated by catalytic activity concentrations of enzymes. Clin. Chem. Lab. Med. **45**(8), 1043–1057 (2007)
19. Farrell, C.L., Nguyen, L., Carter, A.C.: Data mining for age-related TSH reference intervals in adulthood. Clin. Chem. Lab. Med. **55**(10), e213–e215 (2017)
20. Lo Sasso, B., et al.: Reference interval by the indirect approach of serum thyrotropin (TSH) in a Mediterranean adult population and the association with age and gender. Clin. Chem. Lab. Med. **57**(10), 1587–1594 (2019)
21. Mrosewski, I., et al.: Indirectly determined hematology reference intervals for pediatric patients in Berlin and Brandenburg. Clin. Chem. Lab. Med. **60**(3), 408–432 (2021)
22. Zierk, J., et al.: Age- and sex-specific dynamics in 22 hematologic and biochemical analytes from birth to adolescence. Clin. Chem. **61**(7), 964–973 (2015)
23. Zierk, J., Metzler, M., Rauh, M.: Data mining of pediatric reference intervals. J. Lab. Med. **45**(6), 311–317 (2021)

24. Higgins, V., Adeli, K.: Advances in pediatric reference intervals: from discrete to continuous. J. Lab. Prec. Med. **3**(1), 77–82 (2018)
25. Omosule, C.L., et al.: Pediatric ionised calcium reference intervals from archived radiometer data. Clin. Biochem. **104**, 13–18 (2022)
26. Gallo, S., et al.: Redefining normal bone and mineral biochemistry reference intervals for healthy infants in Canada. Clin. Biochem. **47**(15), 27–32 (2014)
27. Gennai, I., et al.: Age- and sex-matched reference curves for serum collagen type I C-telopeptides and bone ALP in children and adolescents: an alternative multivariate statistical analysis approach. Clin. Biochem. **49**(10–11), 802–807 (2016)
28. Monneret, D., et al.: Reference percentiles for paired arterial and venous umbilical cord blood gases: An indirect nonparametric approach. Clin. Biochem. **67**, 40–47 (2019)
29. Ahmed, S., Zierk, J., Khan, A.H.: Establishment of reference intervals for Alkaline phosphatase in Pakistani children using a data mining approach. Lab. Med. **51**(5), 484–490 (2020)
30. Ahmed, S., et al.: Indirect determination of serum creatinine reference intervals in a Pakistani pediatric population using big data analytics. World J. Clin. Pediatr. **10**(4), 72–78 (2021)
31. Ammer, T., et al.: RefineR: a novel algorithm for reference interval estimation from real-world data. Sci. Rep. **11**(1), 16023 (2021)
32. Ha, F., et al.: The reference intervals of whole blood copper, zinc, calcium, magnesium, and iron in infants under 1 year old. Biol. Trace Elem. Res. **200**(1), 1–12 (2022)
33. Dathan-Stumpf, A., et al.: Pediatric reference data of serum lipids and prevalence of dyslipidemia: results from a population-based cohort in Germany. Clin. Biochem. **49**(10–11), 740–749 (2016)
34. Grecu, D.S., et al.: Quality in post-analytical phase: indirect reference intervals for neonates erythrocyte parameters. Clin. Biochem. **46**(7–8), 617–621 (2013)
35. Zierk, J., et al.: Indirect determination of pediatric blood count reference intervals. Clin. Chem. Lab. Med. **51**(4), 863–872 (2013)
36. Weidhofer, C., et al.: Dynamic reference intervals for coagulation parameters from infancy to adolescence. Clin. Chim. Acta **482**, 124–135 (2018)
37. Zierk, J., et al.: Next-generation reference intervals for pediatric hematology. Clin. Chem. Lab. Med. **57**(10), 1595–1607 (2019)
38. Zeljkovic, A., et al.: Indirect reference intervals for haematological parameters in capillary blood of pre-school children. Biochemia Medica **31**(1), 134–142 (2021)
39. Bracho, F.J.: Reference intervals of automated reticulocyte count and immature reticulocyte fraction in a pediatric population. Int. J. Lab. Hematol. **44**(3), 461–467 (2022)
40. Shaw, J.L.V., et al.: Validity of establishing pediatric reference intervals based on hospital patient data: a comparison of the modified Hoffmann approach to CALIPER reference intervals obtained in healthy children. J. Clin. Biochem. **47**(3), 166–172 (2014)
41. Ozarda, Y., et al.: Comparison of reference intervals derived by direct and indirect methods based on compatible datasets obtained in Turkey. Clinica Chimica Acta: Int. J. Clin. Chem. **520**, 186–195 (2021)
42. Haeckel, R.: Indirect approaches to estimate reference intervals. J. Lab. Med. **45**(2), 31–33 (2021)
43. Chaler, E.A., et al.: Age-specific thyroid hormone and thyrotropin reference intervals for a pediatric and adolescent population. Clin. Chem. Lab. Med. **50**(5), 885–890 (2012)
44. Elmlinger, M.W., et al.: Reference intervals from birth to adulthood for serum thyroxine (T4), triiodothyronine (T3), free T3, free T4, thyroxine binding globulin (TBG) and thyrotropin (TSH). Clin. Chem. Lab. Med. **39**(10), 973–979 (2001)
45. Strich, D., Edri, S., Gillis, D.: Current normal values for TSH and FT3 in children are too low: evidence from over 11,000 samples. J. Pediatr. Endocrinol. Metab. **25**(3–4), 245–248 (2012)

46. Aitkenhead, H., Heales, S.J.: Establishment of paediatric age-related reference intervals for serum prolactin to aid in the diagnosis of neurometabolic conditions affecting dopamine metabolism. Ann. Clin. Biochem. **50**(2), 156–158 (2013)
47. Cook, J., et al.: Pediatric reference ranges for prolactin. Clin. Chem. **38**(6), 959 (1992)
48. Roizen, J.D., et al.: Determination of reference intervals for serum total calcium in the vitamin d-Replete pediatric population. J. Clin. Endocrinol. Metab. **98**(12), E1946–E1950 (2013)
49. Imamoglu, E.Y., et al.: Nomogram-based evaluation of thyroid function in appropriate-for-gestational-age neonates in intensive care unit. J. Perinatol. **35**(3), 204–207 (2015)
50. Søeby, K., et al.: Mining of hospital laboratory information systems: a model study defining age- and gender-specific reference intervals and trajectories for plasma creatinine in a pediatric population. Clin. Chem. Lab. Med. **53**(10), 1621–1630 (2015)
51. Den Elzen, W.P.J., et al.: NUMBER: Standardised reference intervals in the Netherlands using a ‘big data’ approach. Clin. Chem. Lab. Med. **57**(1), 42–56 (2018)
52. Zierk, J., et al.: High-resolution pediatric reference intervals for 15 biochemical analytes described using fractional polynomials. Clin. Chem. Lab. Med. **59**(7), 1267–1278 (2021)
53. ÖZYÜREK, E., et al.: Complete blood parameters for healthy, SGA, full-term newborns. Clin. Lab. Haematol. **28**(2), 97–104 (2006)
54. Grecu, D.S., Paulescu, E.: Quality assurance in the laboratory testing process: indirect estimation of the reference intervals for platelet parameters in neonates. Clin. Biochem. **47**(15), 33–37 (2014)
55. Wasiluk, A., et al.: Platelet indices in SGA newborns. Adv. Med. Sci. **56**(2), 361–365 (2011)
56. Toulon, P., et al.: Age dependency for coagulation parameters in paediatric populations. Results of a multicentre study aimed at defining the age-specific reference ranges. Thromb. Haemost. **116**(1), 9–16 (2016)
57. Zierk, J., et al.: Indirect determination of hematology reference intervals in adult patients on Beckman Coulter UniCell DxH 800 and Abbott CELL-DYN Sapphire devices. Clin. Chem. Lab. Med. (CCLM) **57**(5), 730–739 (2019)
58. Herklotz, R., et al.: Metaanalysis of reference values in hematology. Ther. Umsch. **63**(1), 5–24 (2006)
59. Fleming, J.K., et al.: Development of nation-wide reference intervals using an indirect method and harmonised assays. Clin. Biochem. **99**, 20–59 (2022)
60. Sung, J.Y., et al.: Establishment of pediatric reference intervals for routine laboratory tests in korean population: a retrospective multicenter analysis. Ann. Lab. Med. **41**(2), 155 (2021)
61. Griffiths, J.K., et al.: Centile charts II: alternative nonparametric approach for establishing time-specific reference centiles and assessment of the sample size required. Clin. Chem. **50**(5), 907–914 (2004)
62. Koduah, M., Iles, T.C., Nix, B.J.: Centile charts I: new method of assessment for univariate reference intervals. Clin. Chem. **50**(5), 901–906 (2004)
63. Loh, T.P., et al.: Development of paediatric biochemistry centile charts as a complement to laboratory reference intervals. Pathology **46**(4), 336–343 (2014)
64. Arzideh, F., et al.: An improved indirect approach for determining reference limits from intra-laboratory data bases exemplified by concentrations of electrolytes/Ein verbesserter indirekter Ansatz zur Bestimmung von Referenzgrenzen mittels intra-laboratorieller Datensätze am Beispiel von Elektrolyt-Konzentrationen. J. Lab. Med. **33**(2), 52–66 (2009)
65. Arzideh, F., Wosniok, W., Haeckel, R.: Reference limits of plasma and serum creatinine concentrations from intra-laboratory data bases of several German and Italian medical centres: comparison between direct and indirect procedures. Clin. Chim. Acta **411**(3–4), 215–221 (2010)