BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Indirect determination of hemoglobin A2 reference intervals in Pakistani infants using data mining

Muhammad Shariq Shaikh[1*], Sibtain Ahmed[1], Saba Farrukh[2] and Shahnawaz Bayunus[1]

Abstract

Background Reference intervals (RIs) are crucial for distinguishing healthy from sick individuals and vary across age groups. Hemoglobinopathies are common in Pakistan, making the quantification of hemoglobin variants essential for screening. Direct RIs are established by measuring values from a healthy reference population, whereas indirect RIs, use statistical analysis of routine lab data to estimate values, making it feasible in settings where direct data is unavailable. Since Pakistan lacks locally established Hemoglobin A2 RIs for infants, this study aims to fill that gap using an indirect data mining method to improve diagnostic accuracy for hemoglobinopathies.

Methods It was a retrospective observational study. Hemoglobin A2 measurements from all patients aged birth to 1 year between January 2015 and December 2022 were retrieved from the laboratory management system at Aga Khan University Hospital. The study population represented the entire geographical distribution of the country. Hemoglobin A2 was measured using the Bio-Rad Variant™ II analyzer. RIs were computed using an indirect KOSMIC algorithm, which assumes non-pathologic samples follow a Gaussian distribution after Box-Cox transformation.

Results A total of 88,690 specimens were analyzed for HbA2. After excluding patients with multiple specimens, RIs were calculated for 22,713 infants, stratified into five age sub-groups. The 2.5th and 97.5th percentile results showed good agreement with RIs from Mayo Clinic Laboratories.

Conclusions This study supports data mining as an alternative method for establishing HbA2 RIs, especially in resource-limited settings. The results are specific to the studied population, instrument, and reagent, and they elucidate the fluctuations in HbA2 synthesis with age. These intervals will enhance clinical decision-making based on HbA2 results.

Keywords Data mining, Hemoglobin A2, Reference interval, Pakistan, Infants

*Correspondence:
Muhammad Shariq Shaikh
muhammad.shariq@aku.edu
[1]Department of Pathology and Laboratory Medicine, The Aga Khan University Hospital, Stadium Road, Karachi 74800, Pakistan
[2]Clinical Research Fellow, Leeds Teaching Hospitals NHS Trust, Leeds LS97TF, UK

Background

Laboratory investigations continue to be imperative to our understanding of an ever-expanding list of diseases. These have become integral to our diagnosis and line of treatment. Establishing reference intervals (RIs) keeping one's national demographics in mind is in the patient's best interest. Additionally, a comprehensive pathology report allows the healthcare provider to understand the ailment at hand in the best way. The RI is based on the values derived from the distribution of results acquired from a sample reference population. Direct RIs are established by measuring values from a healthy reference population, whereas indirect RIs, use statistical analysis of routine lab data to estimate values, making it feasible in settings where direct data is unavailable. Children often undergo pronounced physiological changes. Therefore, their RIs also differ from the adult cohort [1].

This need for accurate RIs is particularly critical in the context of hemoglobin variants like Hemoglobin A1 (HbA1), Hemoglobin A2 (HbA2), and Fetal Hemoglobin (HbF) that each person possesses. These fractions are crucial for assessing hematological changes in thalassemias and other hemoglobinopathies [2]. Hemoglobin A2 comprises two alpha chains and two delta chains. It is almost absent (0.2–0.3%) at birth and increases in concentration gradually when δ-chain formation begins [3, 4]. In β-thalassemia carriers, the percentage of HbA2 is increased. Therefore, the onus of the diagnosis for thalassemia carriers depends on its accurate quantification [5].

The prevalence of beta-thalassemia varies from 5 to 8% in different regions of Pakistan [6, 7]. An estimated 5000–9000 children are born with beta-thalassemia major each year in the country. For a patient undergoing regular blood transfusions and iron chelation therapy, the annual cost of treatment could range from $4,000 to $18,000 USD.

Given the burden of β-thalassemia in Pakistan, it's vital that RIs of HbA2 fraction be determined and standardized in laboratories. Understandably, most labs are cautious to choose a direct approach right away due to several ethical and practical restrictions. As a developing nation dealing with financial stringency, Pakistan's healthcare system continues to face limitations in carrying out nationwide data collection. So far, western literature has formed the basis of RIs used in Pakistan.

The purpose of this study is to derive HbA2 RIs in Pakistani infants by mining big laboratory data as an indirect approach. Stratified further into narrow age groups, established population-based age-specific RIs will assist more meticulous medical decisions.

Methods

Study design and settings

It was a retrospective observational study. Having received exemption from the institutional ethical review board (2022-7534-21514), the study was carried out at the Section of Hematology and Transfusion Medicine, Department of Pathology and Laboratory Medicine, Aga Khan University Hospital (AKUH), Pakistan. Anonymity was maintained throughout the collection of data; and both inpatient and outpatient data were obtained for infants aged from birth to 1 year. HbA2 measurements were taken from AKUH Clinical Laboratories' laboratory management system from January 2015 to December 2022. With AKU's strong foothold across Pakistan, numerous collection points across Karachi, and other provinces of Pakistan aided us in gathering large data. Strict compliance with predetermined operating procedures was followed through all phases of the testing process. HbA2 is measured in our lab using the BIO-RAD VARIANT™ II Hemoglobin testing system (BIO-RAD Laboratories Inc, Hercules, CA, USA). Two different levels (normal and high abnormal) of commercial control material (Lyphochek® Hemoglobin A2 Control) are used with each batch of samples. To maximize quality assurance, biannual participation in College of American Pathologist surveys take place.

Different age partitions were created to provide a comprehensive representation of the population, ensuring that the RIs are applicable and relevant for clinical practice. Each age group was determined to balance the size of the sample with the need for specificity in the RIs. This approach allows for a more distinct understanding of Hemoglobin A2 levels across different stages of infancy, ultimately supporting better diagnostic accuracy.

Statistical calculation

The RIs were derived for age-distributed subgroups. An indirect algorithm known as KOSMIC was used which has been proposed and validated by Zierk and colleagues [8]. A statistical program which is implemented within a software package to calculate the Box-Cox transformation parameter lambda (λ), the truncation interval, and the parameters of the Gaussian distribution Mu (μ) and sigma (σ) was utilized. This program is based on the proposition that the fraction of physiological samples in the entered dataset can be expressed using parametric distribution. Subsequently, a truncation interval T exists within the dataset in which the percentage of abnormal test results is nominal. Moreover, to project the distribution of non-pathological test results, the lower and upper truncation limits, i.e., T1 and T2, were determined using a "Brute Force" approach. Extensive details of the statistical analysis utilized are available from Zierk J et al. [8]. To

Table 1 Number of HbA2 results analyzed in different age groups

Total results retrieved from LIS	**88,690**
Results analyzed*	22,713
0–59 days	573
60–89 days	859
90–149 days	3260
150–209 days	4991
210–365 days	13,030

* after exclusion of duplicates

support our study, RIs obtained were also compared with those available on Mayo Clinic Laboratories Website [9].

Results

A total of 88,690 specimens for HbA2 were retrieved from LIS during the study period. Excluding patients with > 1 year of age and with > 1 specimen, RIs were calculated for 22,713 infants for HbA2. Results were stratified into 5 age groups; the number of samples analyzed in each age group are shown in Table 1. Figure 1 shows the distribution of physiological test results in each age group. A comparison of our 2.5th and 97.5th percentile results with those of established RIs available at Mayo Clinic Laboratories demonstrated good agreement between different age groups (Table 2).

Table 2 Comparison of Mayo Clinic Laboratories (MCL) and newly established lower (LRI) and Upper Reference intervals (URI) for HbA2 (%)

Age	**New LRI**	**New URI**	**MCL LRI**	**MCL URI**
0–59 days	0.32	2.08	0.0	2.6
60–89 days	0.80	3.08	0.0	2.6
90–149 days	1.75	3.38	1.3	3.1
150–209 days	1.83	3.31	2.0	3.3
210–365 days	1.94	3.4	2.0	3.3

Discussion

The need for reagent and analyzer combination-specific RIs is now well established. The constant introduction of sophisticated instruments and better reagents has further made accurate RIs for all tests a necessity. However, establishing RIs using direct blood sampling is always met with significant challenges. In addition to heavy cost and logistical arrangements, the direct approach requires convincing study participants to give a blood sample voluntarily; the aspect which further becomes complex in the pediatric population. Unless innovative methods that require a minimum blood quantity for reliable measurement are available, ethical dilemmas associated with testing, specifically children, will continue to hinder the establishment of age-specific RIs using a direct approach. By using an indirect data mining approach, these issues

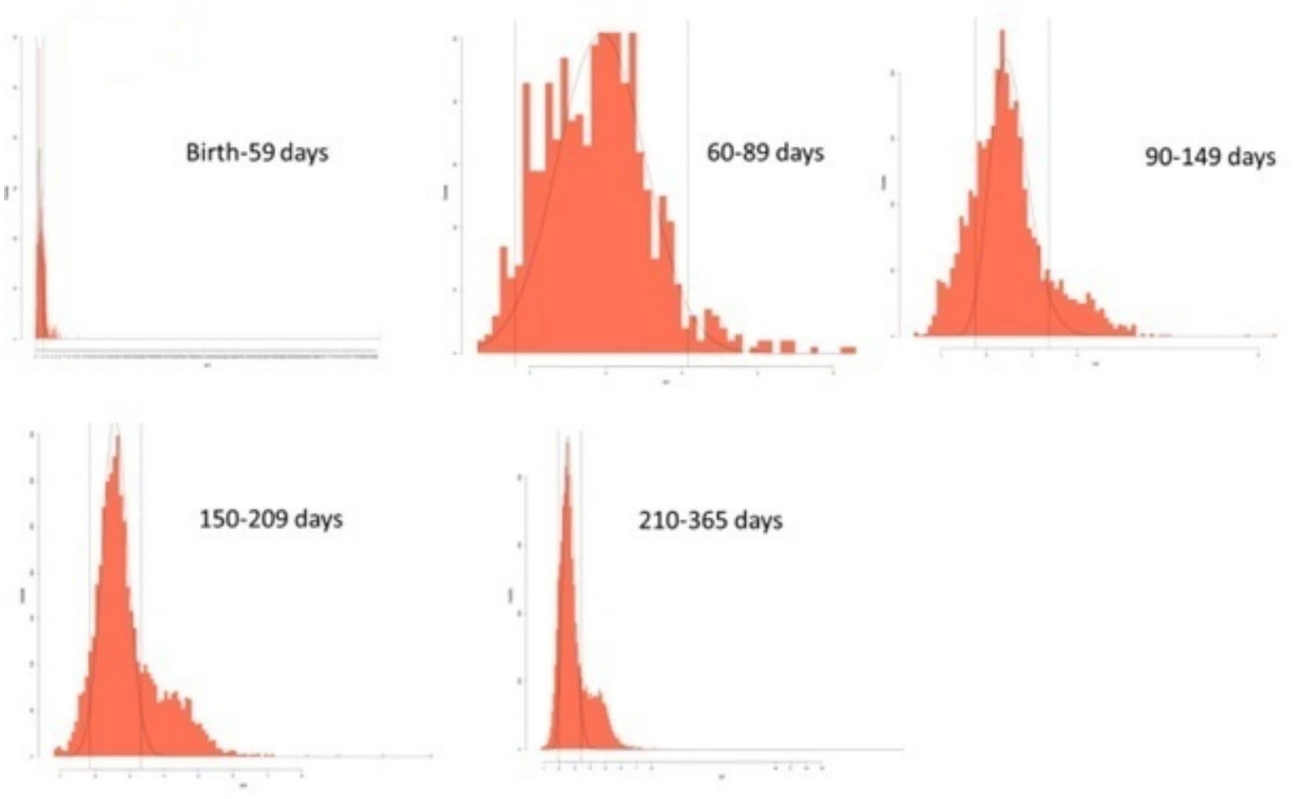

Fig. 1 Distribution of physiological HbA2 test results in different age groups

are overcome as support from IT staff is mainly required for data extraction from the laboratory information system. With a population of over 232 million people, Pakistan stands 5th amongst all countries in the world. This figure is equivalent to 2.83% of the total world population [10]. Despite these noteworthy numbers, population-specific RIs for HbA2 are not yet established using recommended guidelines and appropriate statistical calculations. Current practice by most Pakistani labs is to use published literature as reference in patient reports. Of the current published literature, a majority focuses on data for Western populations. Children are still not a well-studied group in terms of RIs. The reason for the dearth of available published literature is multifactorial. As children undergo drastic physiological changes in early years of life, studies considering these age groups would require complex stratification, hence massive data would be needed [11].

The findings of this study demonstrate strong agreement between the RIs derived through our data mining approach and those established by the Mayo Clinic. However, certain differences were noted between our results and those reported in previous studies. In a study conducted at New Orleans Children's Hospital laboratory, HbA2 range in 65 patients < 5 months old was found to be 0.5–2.9%, whereas it was 1.2–3.2% in children 6 months to 1 year of age [12]. Another study done in Saudi children reported HbA2 level of 3.2 ± 0.2 in 10 children up to 1 year of age [13]. The variance in our study could be attributed to genetic, nutritional, and environmental differences specific to our population. Pakistan's unique demographic makeup and the prevalence of hemoglobinopathies, such as beta-thalassemia, might also contribute to these differences. Lastly, the variation in the testing platform (analyzer, methodology, and reagents) cannot be overlooked as a potential cause of differences in hemoglobin A2 RIs. The approach of determining age-specific RIs is crucial for accurate patient results' interpretation. Our dataset consisted of HbA2 results of 22,713 Pakistani infants, making it the largest sample size of children for RI establishment for any analyte in the country to date. To support our study, RIs obtained using KOSMIC algorithm were also compared with those available on Mayo Clinic Laboratories Website [9]. While the specifics of how these RIs were established are not detailed on their website, they serve as an important benchmark for comparison. Our analysis showed strong agreement between the RIs we derived from a substantial dataset of Pakistani infants and those reported by Mayo Clinic. This correlation underscores the reliability of our results and highlights the necessity of establishing population-specific RIs that accurately reflect the unique demographics and physiological characteristics of the Pakistani population.

We acknowledge that while the KOSMIC algorithm provides RIs, it does not inherently generate confidence intervals (CIs). The absence of CIs is a limitation, as they offer valuable insights into the variability of our estimates. Unfortunately, due to technical constraints, we were unable to apply bootstrap resampling to derive these intervals. This limitation is particularly pertinent given the considerable variation in sample sizes across age groups, which can lead to differing degrees of precision in the RIs. For instance, the younger age group ($n = 573$) may produce wider and less stable intervals compared to the oldest group ($n = 13{,}030$). Therefore, caution is warranted when interpreting the derived RIs. Future research should aim to validate these findings in larger cohorts and explore alternative methods to estimate confidence intervals, such as parametric approaches or additional statistical techniques, to enhance the reliability and applicability of the derived RIs.

Conclusions

We successfully derived age-specific RIs for HbA2 in Pakistani infants using a data mining approach. Our study is the first of its kind to report pediatric HbA2 RIs in the country. By utilizing a large dataset, the intervals determined reflect the HbA2 distribution in the Pakistani infant population. These RIs will enhance clinical decision-making, ensuring accurate diagnosis and better management of hemoglobinopathies in Pakistani children. Future studies should aim to validate these findings in different populations and with varied analytical systems.

Abbreviations

AKUH	Aga Khan University Hospital
CI	Confidence Interval
RIs	Reference Intervals

Acknowledgements

We thank the IT staff at Aga Khan University Hospital for their support in data extraction.

Author contributions

MSS designed the study, collected/analysed data and wrote the initial draft. SA contributed to design, analysis and writing. SF and SB contributed in drafting and data acquisition. All authors critically reviewed and approved the final version of the manuscript.

Funding

Not applicable.

Data availability

The datasets used and analysed during the current study are available from the corresponding author on reasonable request. Most of data analysed during this study are presented in this published article.

Declarations

Ethics approval and consent to participate

The study was approved as exemption and the need for informed consent was waived by Ethics Review Committee of The Aga Khan University (AKU-ERC) under reference number 2022-7534-21514.

Consent for publication
Not applicable as this manuscript contains no individually identifiable details or images.

Competing interests
The authors declare no competing interests.

Received: 28 May 2024 / Accepted: 3 January 2025
Published online: 09 January 2025

References

1. Horn PS, Pesce AJ, Copeland BE. A robust approach to reference interval estimation and evaluation. Clin Chem. 1998;44(3):622–31. https://doi.org/10.1093/clinchem/44.3.622.
2. Weatherall DJ, Clegg JB. Inherited haemoglobin disorders: an increasing global health problem. Bull World Health Organ. 2001;79(8):704–12. https://doi.org/10.1590/S0042-96862001000800011.
3. Cao A, Galanello R, Rosatelli MC, et al. Clinical experience of management of thalassemia: the sardinian experience. Semin Hematol. 1996;33(1):66–75.
4. Kattamis AC, Metaxotou-Mavromati A, Tsiarta H, et al. The course of HbA2 during the first years of life. Acta Haematol. 1975;54(4):221–30. https://doi.org/10.1159/000208866.
5. Bain BJ. Haemoglobinopathy Diagnosis. 2nd ed. Blackwell Publishing; 2006.
6. Khattak MF, Saleem M. Prevalence of heterozygous beta-thalassemia in northern areas of Pakistan. J Pak Med Assoc. 1992;42(2):32–4.
7. Usman M, Moinuddin M. Frequency of β-thalassemia trait in families of patients with β-thalassemia major in Karachi. J Pak Med Assoc. 1998;48(3):70–1.
8. Zierk J, Arzideh F, Haeckel R, et al. KOSMIC: a robust algorithm for outlier detection and reference interval estimation for big data. Clin Chem. 2015;61(10):1206–15. https://doi.org/10.1373/clinchem.2015.240358.
9. Mayo Clinic Laboratories, Test ID. HBA2. [Online] Available: https://www.mayocliniclabs.com/test-catalog/Clinical+and+Interpretive/8288 [Accessed 2024-05-25].
10. Worldometer. Pakistan Population. (2024) - Worldometer. [Online] Available: https://www.worldometers.info/world-population/pakistan-population/ [Accessed 2024-05-25].
11. Soldin SJ, Wong EC, Brugnara C, et al. Pediatric reference intervals. 7th ed. AACC; 2011.
12. Craver RD, Abermanis JG, Warrier RP, et al. Hemoglobin A_2 levels in healthy persons, sickle cell disease, sickle cell trait, and β-thalassemia by capillary isoelectric focusing. Am J Clin Pathol. 1997;107(1):88–91. https://doi.org/10.1093/ajcp/107.1.88.
13. El-Hazmi MA, Warsy AS. Normal reference values for hematological parameters, red cell indices, hb A2 and hb F from early childhood through adolescence in saudis. Ann Saudi Med. 2001;21(3–4):165–9. https://doi.org/10.5144/0256-4947.2001.165.

Publisher's note