Clinica Chimica Acta 561 (2024) 119763

Contents lists available at ScienceDirect

Clinica Chimica Acta

journal homepage: www.elsevier.com/locate/cca

Registered Report Stage II

Use of hospital big data to optimize and personalize laboratory test interpretation with an application

Ronan Boutin [a], Jakez Rolland [a,b], Marie Codet [a], Clément Bézier [a,c,*], Nathalie Maes [d], Philippe Kolh [e], Leila Equinet [a], Marie Thys [f], Michel Moutschen [g], Pierre-Jean Lamy [h,i], Adelin Albert [d,j]

[a] Bio Logbook, 1 rue Julien Videment, 44200 Nantes, France
[b] Nantes University, École Centrale Nantes, CNRS, LS2N, UMR 6004, 1 Rue de la Noë, 44321 Nantes, France
[c] University of Western Brittany, INSERM, LBAI, UMR1227, 9 Rue Félix le Dantec, 29200 Brest, France
[d] Biostatistics and Medico-economic Information Department, University Hospital of Liege, Avenue de l'Hôpital 1, 4000 Liège, Belgium
[e] Department of Information Systems Management, University Hospital of Liege, Avenue de l'Hôpital 1, 4000 Liège, Belgium
[f] Use of medico-economic data, University Hospital of Liege, Avenue de l'Hôpital 1, 4000 Liège, Belgium
[g] Infectious Diseases Department, University Hospital of Liege, Avenue de l'Hôpital 1, 4000 Liège, Belgium
[h] Biopathology and Genetics of Cancers, Institute of Medical Analysis IMAGENOME, INOVIE, 90 rue Nicolas Chedeville, 34075 Montpellier, France
[i] Clinical Research Department, Clinique BeauSoleil, Aesio Santé Méditerranée, 149 Rue de la Taillade, 34070 Montpellier, France
[j] Public Health Department, University of Liege, Avenue de l'Hôpital 1, 4000 Liège, Belgium

ARTICLE INFO

Keywords:
Big data
Optimization
Personalization
Platelets
Precision medicine
Reference population

ABSTRACT

Background and aims: In laboratory medicine, test results are generally interpreted with 95% reference intervals but correlations between laboratory tests are usually ignored. We aimed to use hospital big data to optimize and personalize laboratory data interpretation, focusing on platelet count.
Material and methods: Laboratory tests were extracted from the hospital database and exploited by an algorithmic stepwise procedure. For any given laboratory test Y, an "optimized and personalized reference population" was defined by keeping only patients whose laboratory values for all Y-correlated tests fell within their own usual reference intervals, and by partitioning groups by individual-specific variables like sex and age category. The method was applied to platelet count.
Results: Laboratory data were recorded for 28,082 individuals. At the end of the algorithmic process, seven correlated laboratory tests were chosen, resulting in a reference sample of 159 platelet counts. A new 95 % reference interval was constructed [152–334 $\times$ 10^9/L], notably reduced (27.2 %) compared to conventional reference values [150–400 $\times$ 10^9/L]. The reference interval was validated on a sample of 2,129 patients from another downtown laboratory, emphasizing the potential transference of the hospital-derived reference limits.
Conclusion: This method offers new perspectives in laboratory data interpretation, especially in patient screening and longitudinal follow-up.

1. Introduction

Laboratory medicine has long played a determinant role in the diagnosis, treatment, and surveillance of hospitalized patients. Daily, millions of laboratory tests are performed worldwide and need to be interpreted for clinical decision-making. The most popular way to do this is still the use of reference intervals, a concept developed almost 75 years ago. For most biological parameters, reference intervals were determined by considering a reference population of presumably healthy individuals and including 95 % of their values. For Gaussian

Abbreviations: ANOVA, Analysis of Variance; CLSI, Clinical & Laboratory Standards Institute; CRP, C Reactive Protein; MCHC, Mean Corpuscular Hemoglobin Concentration; MCV, Mean Corpuscular Volume; MHC, Mean Corpuscular Hemoglobin; RBC, Red Blood Cell; RDI, Relative Dispersion Index; WBC, White Blood Cell.

* Corresponding author.
E-mail addresses: ronan.boutin@biologbook.fr (R. Boutin), jakez.rolland@biologbook.fr (J. Rolland), marie.codet@biologbook.fr (M. Codet), clement.bezier@biologbook.fr (C. Bézier), nmaes@chuliege.be (N. Maes), philippe.kolh@chuliege.be (P. Kolh), leila.equinet@biologbook.fr (L. Equinet), mthys@chuliege.be (M. Thys), mmoutschen@chuliege.be (M. Moutschen), pierre-jean.lamy@labosud.fr (P.-J. Lamy), aalbert@uliege.be (A. Albert).

https://doi.org/10.1016/j.cca.2024.119763
Received 27 October 2023; Received in revised form 29 April 2024; Accepted 3 June 2024
Available online 6 June 2024

distributions, the lower and upper limits are obtained by calculating the mean ± 2 standard deviations of a sample of at least 120 healthy subjects. For skewed distributions that cannot be normalized, the lower and upper limits are defined by the 2.5th and 97.5th percentiles, respectively. Then, any test result outside the reference limits is considered "pathological or abnormal", and "normal" otherwise [1,2]. Subsequently, 95 % reference intervals were stratified by sex, age class, ethnicity [3–8], blood group [9] or given some genetic information [10]. Some authors even proposed individual reference intervals based on data accumulated over time in healthy subjects, for example in people undergoing systematic biological check-ups [11].

The closeness of the reference population to a perfectly healthy population depends on the inclusion and exclusion criteria used to define it and to reduce between-subjects variability [12]. Basically, the concept of a healthy population is purely hypothetical because some unsuspected clinical or laboratory features cannot be fully excluded, even when people are apparently healthy and free of any serious condition [13–16]. It is also known that biological parameters are often correlated, so that defining a reference interval without accounting for these correlations may not be optimal [17]. Thus, including the biological variations of all correlated parameters, or at least as many as possible, in the selection of reference individuals may lead to a novel vision of the reference population concept [17,18].

Hospital laboratories generate daily vast quantities of test results stored in data warehouses which may be exploited to refine laboratory tests interpretation. There have been attempts in the past to develop reference intervals from patient data but without real success [19]. Nonetheless there is a potential in hospital laboratory data because they come from diseased patients as well as from healthy individuals. Therefore, not all results are abnormal. Complex statistical algorithms are now available to discard data from non-healthy individuals [14]. They can also identify biological parameters associated to the one under study and keeping only those individuals for whom all biological parameters results are normal with respect to their usual reference intervals [18]. The present research work purposed to determine refined reference populations and corresponding 95 % reference intervals from hospital big data in the context of precision and personalized medicine. The method is illustrated for platelet count.

2. Materials and methods

2.1. Patient databases

This study was based on retrospective data from a hospital laboratory (Liege University Hospital) and a downtown laboratory (Montpellier, France). This included 28,356 distinct patients aged ≥ 18 years and hospitalized between 2005 and 2019 in the University Hospital (Liege, Belgium), yielding a total of 1,692,564 laboratory test results. The patient mean age was 56.1 ± 16.9 years (range 18–98 years), and there were 47.4 % of males and 52.6 % of females. For each subject, the laboratory report comprised at least red blood cell (RBC) count, hemoglobin, hematocrit, mean corpuscular volume (MCV), mean corpuscular hemoglobin (MHC), mean corpuscular hemoglobin concentration (MCHC), platelet count, white blood cell (WBC) count (total leucocytes, lymphocytes, monocytes, neutrophils, eosinophils, and basophils). This constituted the hospital training dataset. The validation dataset was drawn from a downtown laboratory database (Montpellier, France), including all subjects aged ≥ 18 years which were tested between October 2016 and December 2019. It consisted of 996,975 distinct individuals (42.2 % of males and 57.8 % of females) with a mean age of 54.5 ± 19.9 years (range 18–110 years) and 37,677,310 test results. As for the training dataset, the laboratory report of each patient included at least the same hematological parameters listed above. The sex considered in this study is that known by the clinical laboratories (Liège University Hospital laboratory and Montpellier downtown laboratory) in its own informatic system.

We focused our study on platelets, lymphocytes, alpha-1 globulins, gamma globulins, blood glucose, calcium, CRP, and 25-hydroxyvitamin D. In Liege University Hospital laboratory, hemograms were performed on 3.0 mL EDTA tubes and analyzed using ADVIA 120 and 2120 (Siemens, Erlangen, Germany) and XE Analyzer (Sysmex, Kobe, Japan). Protein electrophoresis was carried out on 3.5 mL serum (dry tube) and analyzed using capillary electrophoresis by V8 Nexus CE (HELENA Laboratories, Beaumont, U.S.A) from 2005 to 2011 and CAPILLARYS 3 (Sebia, Lisses, France) from 2011 to 2019. Fraction measurements, initially recorded in percentages, were converted to g/L by multiplying with the total serum protein concentration. Glycemia data, collected from both fasting and non-fasting patients, were obtained from either 5.0 mL fluoride or 5.0 mL heparin tubes. Analysis was performed using COBAS Integra 400 plus and Modular analyzer from 2005 to 2012, and COBAS 8000 and COBAS 6000 analyzer from 2012 to 2019 (Roche diagnostics, Bâle, Suisse). Calcium levels were measured in serum (dry tube) or 5 mL heparinized tubes using the Arsenazo III method by the same systems. The C Reactive Protein (non-ultrasensitive assay) was quantified in serum (dry tube) using the same systems through immunoturbidimetry. 25-hydroxyvitamin D levels in serum (dry tube) were determined using a competitive chemiluminescence immunoassay (CLIA) by LIAISON XL (Diasorin, Saluggia, Italy). Graphically, no difference was observed in the values of the studied biological parameters over the years, except for calcium, which exhibited a change in 2012. The difference between values before and after 2012 (92.82 and 94.85 mg/L respectively, 2.07 mg/L in difference) was lower than the within-subject coefficient of variation (CVi) of calcium (18.10 mg/L) determined in the European Biological Variation Study (EuBIVAS) in 2018 on 91 healthy European volunteers [20].

In the downtown laboratory, hemograms were performed on 3.0 mL EDTA tubes and analyzed using DxH 900 with CellaVision DM9600 (Beckman Coulter Inc., Brea, USA). Protein electrophoresis was carried out on 3.5 mL serum (dry tube) and analyzed using capillary electrophoresis CAPILLARYS 3 TERA (Sebia, Lisses, France). Fraction measurements, initially recorded in percentages, were converted to g/L by multiplying with the total serum protein concentration. Glycemia data, collected from both fasting and non-fasting patients, were obtained from either 5.0 mL fluoride or 5.0 mL heparin tubes. Calcium was collected in serum (dry tube). Both glycemia and calcium were measured by spectrophotometric method. Finally, C Reactive Protein (non-ultrasensitive assay) was quantified in serum (dry tube) by immunoturbidimetry method. Glycemia, calcium and CRP were analyzed by COBAS 8000 analyzer (Roche diagnostics, Bâle, Suisse).

2.2. Statistical methodology

2.2.1. The definition of optimized and personalized reference population

Hereafter, let Y denote the laboratory test for which an optimized and personalized 95 % reference interval will be determined. Further, let X denote another laboratory test or subject-specific characteristic (sex, age for example) consisting of several categories, say $C_1, \ldots, C_m$. For sex, $m = 2$ but for age there may be more categories. For a quantitative laboratory test, X will be split into $m = 3$ categories depending on whether the result falls below (C_1), within (C_2) or above (C_3) the 95 % reference interval used in the laboratory. In general, there will be a whole series of covariates (biological parameter and subject-specific characteristics) distinct from Y. The basic idea of deriving an optimized and personalized reference population for Y can be stated as follows. From a collection of N individuals with n covariates X (biological parameters and/or subject-specific characteristics), the reference individuals retained will be those for whom all Y-correlated laboratory parameters values fall within their corresponding 95 % reference intervals (C_2). If an individual undergoes multiple compliant test panels on different dates, only one reference value per individual is selected, with preference given to the test panel with the fewest tests performed on the same date. The hypothesis is that healthy individuals have fewer tests

prescribed than others.

Then, these individuals will be further partitioned by subject-specific characteristics categories. Ultimately, there will be N* (≤N) such "reference" individuals for any subgroup and a 95 % reference interval can then be determined in the classical way (parametric/nonparametric) for optimized and personalized interpretation of *Y*. The benefit of this approach can be assessed by the Relative Dispersion Index (RDI) defined as the ratio of *L* (the width of the laboratory reference interval) and *L** (the width of the optimized and personalized reference interval), specifically RDI = *L/L**. Thus, the new 95 % reference interval for *Y* will be optimized if RDI > 1.

2.2.2. Algorithmic process

An algorithmic stepwise iterative process was developed to find all covariates *X* related to *Y* and determine optimized and personalized reference populations.

2.2.2.1. Step 1. The first step consists in finding all subject-specific characteristics and biological parameters *X* which are individually significantly associated with *Y*. This can be done by one-way ANOVA or equivalently by selecting a "reference category" (say C_1) and applying a multiple regression analysis of *Y* on the other categories C_2, …, C_m of *X*. Note that the quality of the association assessed by R^2, the coefficient of determination, is independent of the choice of the reference category; moreover, the closer R^2 to 1, the stronger the association. A covariate *X* will be selected and declared significantly related to *Y* if the overall p-value of the ANOVA is less than 5 % ($p < 0.05$) or if any category of *X* leads to a significant p-value. In general, there will be many *X* selected in the first step, which we denote $X = X_1$, each yielding a potential model of the form $Y = f(X_1)$. Let M_1 be the set of such models, on which the reference individual selection process described above is applied, allowing to create optimized reference populations with individuals in category C_2 for X_1. Only the 50 models with the largest R^2 values and the 50 models with the greatest RDI will be retained. Let M_1' the subset of these 100 models.

2.2.2.2. Step 2. For each model in M_1' (selected in step 1), a two-way ANOVA will be applied to find any biological parameters and subject-specific characteristics *X* , different from X_1, which combined to X_1 is significantly associated with *Y*. This is like applying a multiple regression analysis of *Y* on any two variables X_1 and *X* , where X_1 is the variable selected in step 1. Here again the quality of the regression of *Y* on the two variables can be assessed by R^2, the global p-value and the p-value of all categories of the two variables. This leads anew to a large set of models in the form $Y = f(X_1, X_2)$, denoted by M_2. As before, models in M_2 allows to create optimized reference populations with individuals in category C_2 for X_1 and X_2. As in step 1, M_2' is composed of the 50 models of M_2 with the largest R^2 values and the 50 models of M_2 with the greatest RDI scores.

2.2.2.3. Following steps and stopping criterion. The procedure described in step 2 is pursued by applying multiway ANOVA in step 3 with 3 variables included. For each model in M_2', each covariate *X* (not already in the model) is tested. Again, a set of models M_3 in the form $Y = f(X_1, X_2, X_3)$ is obtained, and can be reduced to M_3'. The process continues with 4, 5, …, n variables *X* , until no improvement can be added to the last variable tested. We set a maximum at n = 10. Note that multiway ANOVA decomposes the total variability of *Y* into the variability explained by each covariates *X* and the residual variability. In the end of the process, we can consider the set of all models $M = M_1 \cup M_2 \cup \ldots \cup M_n$. A flowchart of the statistical method is provided (see Supplementary Material).

2.3. Validation and transference

To validate this approach, a bootstrap validation was carried out to validate the transference of the hospital-derived reference intervals to the validation dataset. The models were applied to the validation dataset: the reference individuals selected (from the validation dataset) were those for which all the values of the laboratory parameters correlated to Y (according to the model tested) fell within their corresponding 95 % reference intervals, called validation reference individuals. In each 1000 bootstrap of 120 individuals out of the total, the rate of validation reference individuals excluded by the hospital-derived reference interval was calculated. According to the validation test proposed in the 2010 EP28 A3C guideline by the Clinical & Laboratory Standards Institute (CLSI) [1], the transference was validated if no more than 10 % of the test individuals was excluded by the new reference interval under study.

This work has been drawn up in compliance with the Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals (ICMJE) [21] and with The Code of Ethics of the World Medical Association (Declaration of Helsinki). This article was written in compliance with the STROBE Statements for observational studies [22], and SAGER guidelines for Sex and Gender Equity in Research [23].

3. Results

The above-described procedure was applied to more than 100 biological parameters and for each of them hundreds of models were developed and were potential candidates for defining optimized and personalized reference intervals to enhance daily laboratory data interpretation. As an illustration, the procedure will be described for platelet count.

3.1. Optimized reference intervals

Based on the 28,356 patients of the hospital database, the application of the algorithmic procedure yielded 32,927 potential models for platelet count and 385,367 reference populations were determined, of which 93,117 with N*>120 and 19,342 with N*>120 and RDI > 1. Among these, the one with the greatest RDI score (RDI = 1.37) evidenced 7 biological parameters jointly correlated with platelet count: α1-globulins [95 % reference interval: 2.1–3.5 g/L], calcium [83–102 mg/L], C-Reactive Protein (CRP) [0–5 mg/L], γ-globulins [8–13.5 g/L], glycemia [0.6–1.0 g/L], lymphocyte count/total leucocytes [20—40 %], and 25-hydroxyvitamin D [30–60 ng/mL] (Table 1). The association was particularly strong with α1-globulins, calcium, and lymphocytes. Univariately, platelet counts differed significantly according to each selected biological parameter category (below, within or above their 95 % reference interval), except for glycemia (Fig. 1). However, glycemia had a significant effect on platelet count in multi-way ANOVA, thus

Table 1
Joint effect of significant biological parameters with platelet count as assessed by multi-way analysis of variance.

Parameter	Df	SSQ[a]	MSQ[a]	F test[a]	p-Value
Alpha-1 globulins	2	8.55	4.27	40.01	<0.0001
Calcium	2	3.30	1.64	15.44	<0.0001
C Reactive Protein	2	0.45	0.45	4.25	0.039
Gamma globulins	2	1.68	0.84	7.89	0.0.00038
Glycemia	2	1.09	0.54	5.09	0.0062
Lymphocyte count/total leucocytes	2	5.04	2.52	23.59	<0.0001
25-hydroxyvitamin D	2	0.98	0.49	4.58	0.010
Residuals	3612	385.87	0.107		

[a] Multiplied by 10^{11}; SSQ sum of squares (related to parameter); MSQ mean square = SSQ/Df; Df degrees of freedom; F test given by Parameter MSQ /Residuals MSQ with 2 and 2990 degrees of freedom.

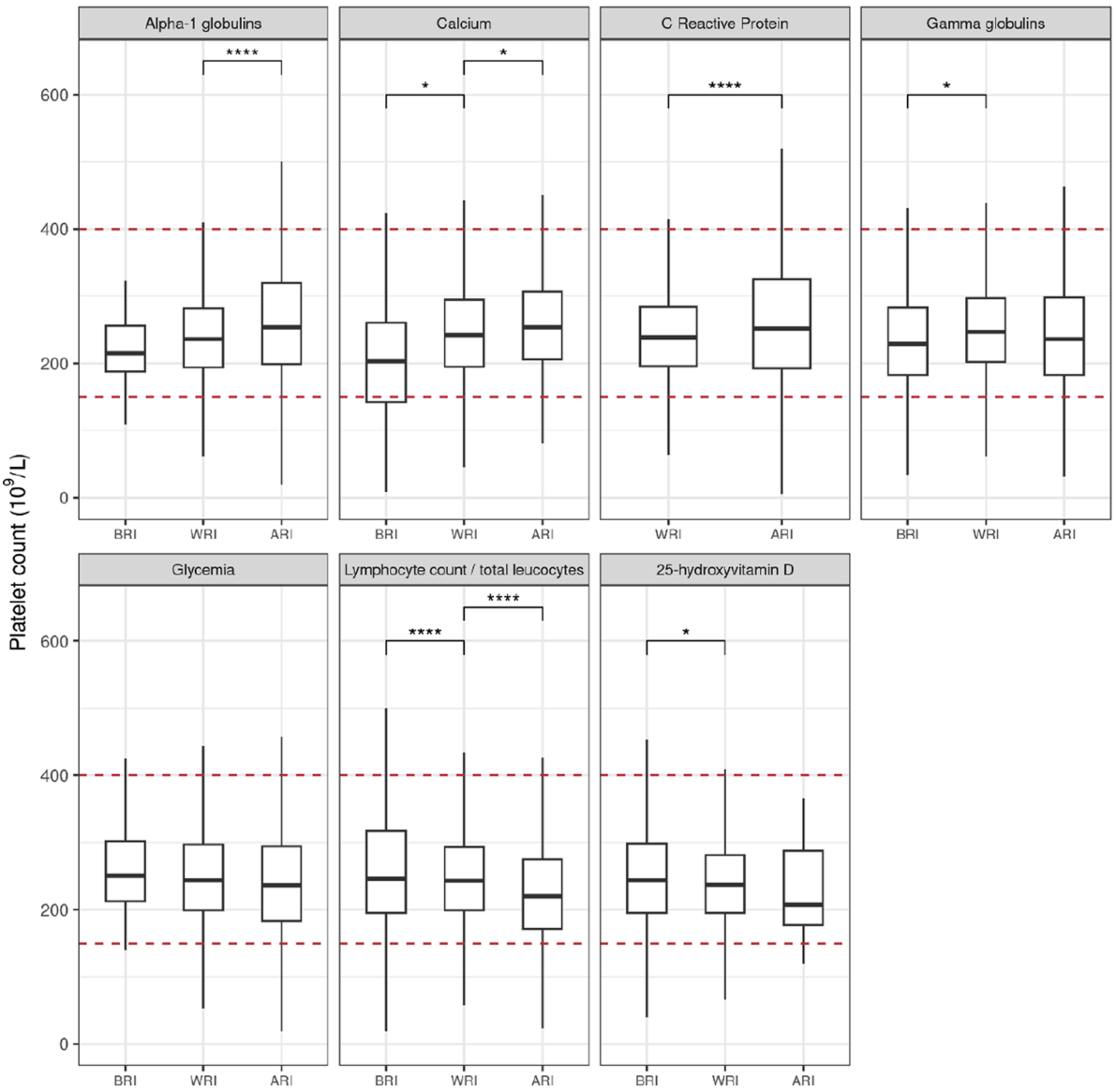

Fig. 1. Platelet count displayed against each correlated biological parameter selected in the model. Correlated parameters are categorized in BRI (Below the 95 % laboratory Reference Interval), WRI (Within the 95 % laboratory Reference Interval) and ARI (Above the 95 % laboratory Reference Interval). Red dotted horizontal lines represent current laboratory 95 % reference interval for platelet count. Significance indicated by asterisks were obtained by t-tests between categories (*: p-value < 0.05, **: p-value < 0.01, ****: p-value < 0.0001).

explaining its presence in the model (Table 1). Formally, the final model was written "Platelets $= f(\alpha_1$-globulins; calcium, CRP, γ-globulins; glycemia; lymphocytes; 25-hydroxyvitamin D)" ($p < 0.0001$). applying these biological exclusion criteria (see section 2.2.1), N*=159 individuals were selected. The distribution of platelet counts in these subjects is displayed together with the reference limits of the hospital laboratory [150–400 $\times$ 10^9/L] (Fig. 2). The refined 95 % reference interval was reduced by 27.2 % and turned out to be [152–334 $\times$ 10^9/L]. While the lower limits were almost superimposable, the upper limits differed markedly.

3.2. Validation and transference

By applying the same model on the validation dataset, N*=2,129 validation reference individuals were retained (1066 females, 1063 males after balancing). The distribution of platelet count of these subjects is also displayed in with a substantial overlap with the one derived from the training dataset (Fig. 2).

In the 1000 bootstraps of 120 individuals out of the N*2,129, on average 4.7 (3.9 %) were below the hospital-derived interval, 7.1 (5.9 %) were above, resulting in 11.8 (9.8 %) validation reference individuals being excluded by the hospital-derived interval. This proportion of 9.8 % out-of-range individuals was lower than the pre-

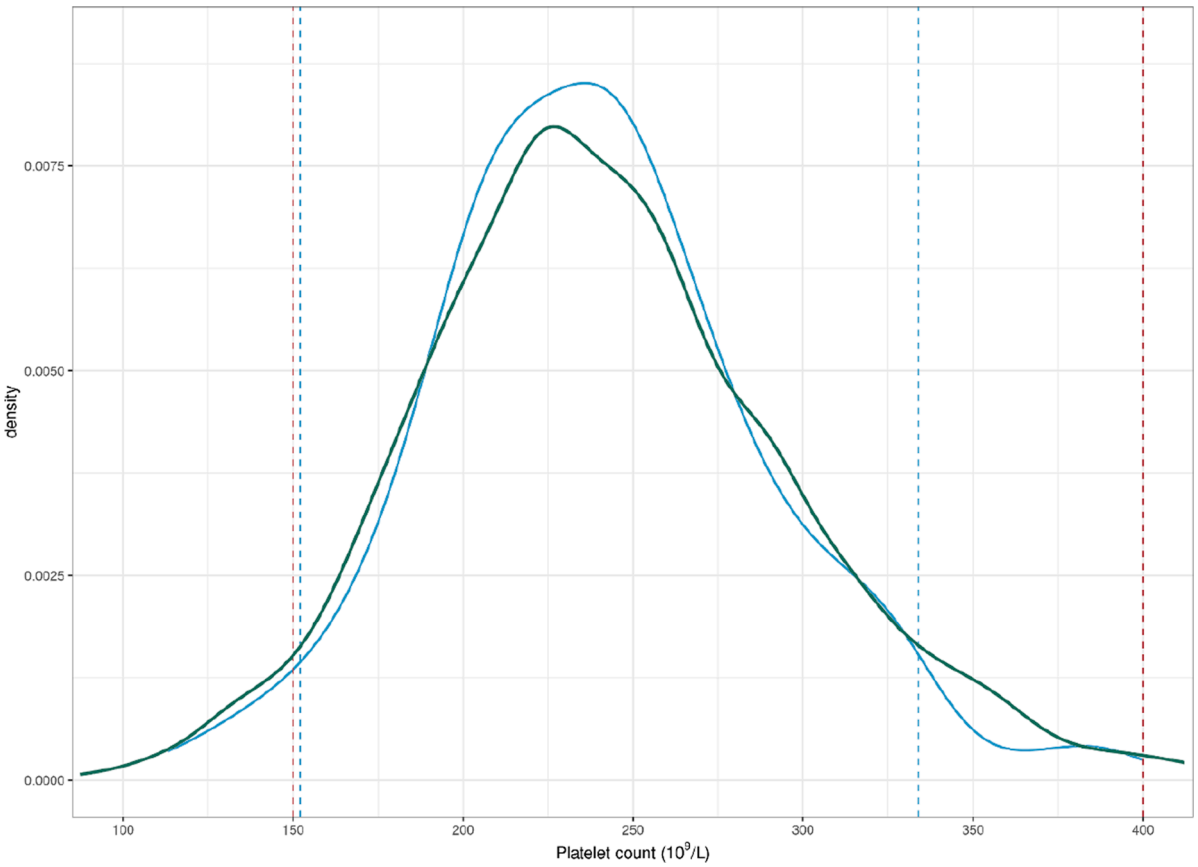

Fig. 2. Density distribution of platelet count (x10^9/L) in the optimized reference population from hospital training data sample (in blue) and in optimized reference population from downtown laboratory validation data sample (in green). The optimized 95% reference interval from hospital data sample is indicated by the blue vertical dotted lines whereas the current 95% reference interval is represented by the red vertical dotted lines.

established threshold of 10 %, thus sustaining the clinical relevance and potential transference of the reference interval determined from hospital big data to the outside laboratory.

4. Discussion

When two laboratory parameters are correlated, their 95 % reference intervals should account for that correlation. This is challenging because the problem becomes rapidly intractable as the number of correlated parameters increases. One way to circumvent this problem is to proceed as with sex or age categories, namely, to split the correlated biological parameters into categories and define 95 % reference intervals of the other laboratory test for each category. We used hospital big data because not all hospital laboratory results are abnormal in the classical way and not all hospitalized patients are diseased. For a given laboratory test (e.g., platelet count), our approach searched for the best model composed of correlated biological and subject-specific parameters that were directly associated with the laboratory test under study. By considering only individuals with values of the correlated parameters within their respective 95 % reference intervals, this leaves for the laboratory test under study N* individuals who define an optimized reference population from which an optimized and personalized 95 % reference interval can be constructed. If this interval is shorter than the 95 % reference interval used in the laboratory, a substantial gain in data interpretation can be obtained. For instance, early detection of variation in a given patient may be uncovered, thus improving diagnosis and follow-up [11], as a complement of routine practice based primarily on symptom onset [24]. The rationale is that, if a laboratory value is within the classical 95 % reference interval used in the laboratory but outside the optimized 95 % reference interval, the value of another correlated biological parameter could be outside its own 95 % reference interval. Our approach also allows calculating mean values that can be considered as the "personalized optimum" for a given laboratory test.

The method was tested on platelet count hospital data. From the total number of patients, N*=159 individuals were ultimately retained by the algorithmic procedure, a sample size high enough to determine a 95 % reference interval. The upper limit (334 × 10^9/L) of the interval derived was 66 units lower than the 400 × 10^9/L conventional limit or lower than the 350 × 10^9/L threshold sometimes used for thrombocytosis diagnosis [25–27]. This is consistent with a Canadian study [28] which determined 95 % reference intervals for platelet count by the direct *a priori* method, as recommended by the CLSI C28 A3 guideline [1]. Their 95 % reference intervals for subjects aged 27–79 years were [151.8–324.0 × 10^9/L] for males and [153.2–361.3 × 10^9/L] for females, respectively [28]. These values are strikingly close to ours and kind of comfort the idea that hospital big data bears potential information when appropriately processed. It has been reported that platelet values may be influenced by age, sex [7,29], geography [29], ethnicity [7,8], or genetic factors [30]. Similar approaches based on systematic biological parameter abnormality elimination have been developed before [18]. In our study, platelet values were primarily correlated with seven biological parameters. The association between platelet count and 25-hydroxyvitamin D was previously reported in another study [31]. As expected, platelet count was also correlated with inflammatory biomarkers (CRP, alpha-1 and gamma globulins, calcium) and with lymphocytes. In other models (not reported here), platelet count was found correlated with mean corpuscular volume, total leucocytes, α2-globulins, cholesterol, and creatinine, which shows that other models can be built to refine and have a more precise reading of platelet results. Age was often related to platelet count, but in the present case partitioning

was not possible due to lack of data. By contrast, sex was not always significantly correlated with platelets, unlike in other studies [7,8,28,29]. It is likely that the absence of sex may result from the fact that by considering the biological parameters correlated with platelets, the effect of sex vanished.

The gain of dispersion for the upper limit of the 95 % optimized reference interval of platelet count may improve identifying early thrombocytosis emergence, although no change was observed for the lower limit and hence for diagnosing thrombocytopenia (platelet count $< 150 \times 10^9$/L). Besides their role in hemostasis, platelets are implicated in inflammation and tumorigenesis [32,33]. Thrombocytosis is widely associated with poor prognosis in non-small cell lung cancer, gastric, pancreatic cancer and others [34–36]. Therefore, it is thought that optimized reference intervals may contribute to better prognosis in these cancer types. Lastly, we believe that our approach based on big data may help creating specific reference geriatric populations (i.e., presumably healthy subjects aged > 80 years), which remains a challenging issue today [2], and discerning "normal" or physiological aging trajectories from an actual diseased state in elderly.

Within the limits of our study, genetic, and environmental factors were not envisaged. About blood glucose, it would have been interesting to analyze blood glucose levels of fasting patients only, but the data from the University Hospital of Liège did not allow us to distinguish between fasting and non-fasting individuals. For the sake of comparability between the two databases, we have included both fasting and non-fasting glucose values in the validation dataset (from the downtown laboratory) too. In both the training and validation datasets, we utilized the fasting blood glucose reference range [0.6–1.0 g/L] as a biological exclusion criterion. It is likely that non-fasting individuals have a blood glucose level exceeding 1.0 g/L and have been naturally excluded from the optimized reference population for platelets (see section 2.2.1). Considering 25-hydroxyvitamin D, data were collected throughout the years. 25-hydroxyvitamin D exhibits significant variability based on the season. We aimed to exclude all individuals with deficiencies (<30 ng/mL) from the optimized reference population both in winter and summer. We visually and statistically verified that the values of 25-hydroxyvitamin D and platelet count in the optimized reference population did not differ based on the time of year (October to March and April to September). P-value = 0.568 and 0.549, for 25-hydroxyvitamin D and Platelet count, respectively. In conclusion, the process of selecting reference individuals (Section 2.2.1) effectively limits the variability of 25-hydroxyvitamin D throughout the year. Finaly, while optimized 95 % reference interval for platelet count established using hospital big data can be generalized to Caucasian populations of Western countries [7], their extension to other populations (e.g., African, Chinese or Iranian) might be difficult [37–39].

The present method is a new indirect methodology for determining 95 % reference intervals of laboratory tests. Even if indirect methods allow the collection of large patient samples, cost and resource savings, direct approaches are recommended by the 2010 EP28 A3C guideline by CLSI [1], and many large cohort studies have followed this principle [3–6,28]. However, ignoring other biological parameters in the reference individual selection process may lead to increased variability in biological profiles. The present method, by the strict reference individual selecting process and the consideration of subject-specific characteristics (sex and age category for example), can further reduce the inter-individual variability in reference populations.

The primary objective of this work was not to define a perfectly or presumably "healthy" population but to obtain homogeneous biological profiles. If we cannot replace the existing 95 % reference intervals used in laboratory daily practice, our method allows a more refined and precise reading of laboratory tests within these reference intervals. Secondly, the same individual can be excluded from a reference population for a given laboratory test and included in the reference population for another laboratory test. This leads to reconsider the idea that an individual is either healthy or diseased.

In the future, other subject-specific characteristics, such as height, weight, blood groups, infectious status, genetics, geographical origin, could be investigated and our approach could be generalized on omics data. We aim in the future at defining multi-dimensional optimized and reference regions capturing all parameters correlations based on innovative approaches in data-science [40]. The challenge is now to define reference populations that integrate these functionalities on an appropriate platform for use by clinicians, laboratory medicine specialists, and researchers.

5. Conclusion

In conclusion, our study introduces a novel indirect methodology for constructing optimized and personalized reference populations, allowing for a refined interpretation of laboratory test results. By incorporating correlated parameters into the exclusion criteria for reference individuals, we successfully reduced inter-individual variability within these populations, leading to the establishment of finer reference intervals than those used today in clinical practice. Our model, which considered α1-globulins, calcium, C-Reactive Protein (CRP), γ-globulins, glycemia, lymphocyte count/total leucocytes, and 25-hydroxyvitamin D, provided a tailored reference population for platelet counts. This, in turn, enabled the determination of a new thrombocytosis threshold, potentially offering diagnostic and prognostic advantages. The transferability of this reference population has been validated on an independent data sample, drawn from a downtown laboratory. This suggests that our indirect method can be used to determine new reference intervals on hospital data.

Research ethics

This retrospective study has obtained the approval of University Hospital Ethics Committee of Liege (707) before data analysis and patient consents were respected.

Informed consent

Informed consent was obtained from all individuals included in this study, or their legal guardians or wards.

Author contributions

All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Research funding

None declared.

Data availability

Not applicable.

CRediT authorship contribution statement

Ronan Boutin: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization. **Jakez Rolland:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Marie Codet:** Data curation, Formal analysis, Investigation, Methodology, Software, Validation. **Clément Bézier:** Conceptualization, Data curation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing. **Nathalie Maes:** Data curation, Project administration, Resources, Software, Validation. **Philippe Kolh:** Project administration, Resources, Software, Supervision, Validation. **Leila Equinet:** Conceptualization, Investigation, Project administration, Resources, Supervision, Validation. **Marie Thys:** Data curation, Project administration, Resources, Software, Validation. **Michel Moutschen:** Project administration, Resources, Supervision, Validation. **Pierre-Jean Lamy:** Formal analysis, Investigation, Methodology, Resources, Supervision, Validation, Writing – original draft. **Adelin Albert:** Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank the University Hospital of Liege for its collaboration in the project and the access to the database. We would also like to thank all the patients who gave their consent for their data to be used in this study.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cca.2024.119763.

References

[1] Clinical and Laboratory Standards Institute, editor. Defining, establishing and verifying reference intervals in the clinical laboratory: approved guideline. 3. https://clsi.org/standards/products/method-evaluation/documents/ep28/, 2016 (accessed 18 April 2024).
[2] Y. Ozarda, V. Higgins, K. Adeli, Verification of reference intervals in routine clinical laboratories: practical challenges and recommendations, Clin. Chem. Lab. Med. 57 (2019) 30–37, https://doi.org/10.1515/cclm-2018-0059.
[3] M.F. Strand, P.M. Fredriksen, M. Lindberg, Hematology reference intervals in 6–12-year-old children: the health-oriented pedagogical project (HOPP), Scand. J. Clin. Lab. Investig. 82 (2022) 404–409, https://doi.org/10.1080/00365513.2022.2100820.
[4] M.K. Bohn, V. Higgins, H. Tahmasebi, A. Hall, E. Liu, K. Adeli, et al., Complex biological patterns of hematology parameters in childhood necessitating age- and sex-specific reference intervals for evidence-based clinical interpretation, Int. J. Lab. Hematol. 42 (2020) 750–760, https://doi.org/10.1111/ijlh.13306.
[5] H. Tahmasebi, V. Higgins, M.K. Bohn, A. Hall, K. Adeli, CALIPER Hematology Reference Standards (I): Improving Laboratory Test Interpretation in Children (Beckman Coulter DxH 900–Core Laboratory Hematology System), Am. J. Clin. Pathol. 154 (2020) 330–341, https://doi.org/10.1093/ajcp/aqaa059.
[6] V. Higgins, H. Tahmasebi, M.K. Bohn, A. Hall, K. Adeli, CALIPER Hematology Reference Standards (II): Improving Laboratory Test Interpretation in Children (Beckman Coulter DxH 520–Physician Office Hematology System) With Analytical Comparison to the Beckman Coulter DxH 900, Am. J. Clin. Pathol. 154 (2020) 342–352, https://doi.org/10.1093/ajcp/aqaa057.
[7] J.B. Segal, A.R. Moliterno, Platelet counts differ by sex, ethnicity, and age in the United States, Ann. Epidemiol. 16 (2006) 123–130, https://doi.org/10.1016/j.annepidem.2005.06.052.
[8] B.J. Bain, Ethnic and sex differences in the total and differential white cell count and platelet count, J. Clin. Pathol. 49 (1996) 664–666, https://doi.org/10.1136/jcp.49.8.664.
[9] Z. Chen, X. Dai, J. Cao, X. Tan, S. Chen, M. Yu, Reference intervals for coagulation tests in adults with different ABO blood types, J. Clin. Lab. Anal. 36 (2022) e24269.
[10] Y. Ozarda, Reference intervals: current status, recent developments and future considerations, Biochem. Med. 26 (2016) 5–11. 10.11613/BM.2016.001.
[11] M. Pusparum, G. Ertaylan, O. Thas, Individual reference intervals for personalised interpretation of clinical and metabolomics measurements, J. Biomed Inform. 131 (2022) 1532, https://doi.org/10.1016/j.jbi.2022.104111.
[12] F. Ceriotti, R. Hinzmann, M. Panteghini, Reference intervals: the way forward, Ann. Clin. Biochem. 465 (2009) 8–17, https://doi.org/10.1258/acb.2008.008170.
[13] R.F. Ritchie, G. Palomaki, Selecting clinically relevant populations for reference intervals, Clin. Chem. Lab. Med. 42 (2004) 702–709, https://doi.org/10.1515/CCLM.2004.120.
[14] A. Katayev, C. Balciza, D.W. Seccombe, Establishing reference intervals for clinical laboratory test results: is there a better way? Am. J. Clin. Pathol. 133 (2010) 180–186, https://doi.org/10.1309/AJCPN5BMTSF1CDYP.
[15] R. Gräsbeck, The evolution of the reference value concept, Clin. Chem. Lab. Med. 42 (2004) 692–697, https://doi.org/10.1515/CCLM.2004.118.
[16] C. Petitclerc, Normality: the unreachable star? Clin. Chem. Lab. Med. 42 (2004) 698–701, https://doi.org/10.1515/CCLM.2004.119.
[17] V. Higgins, S. Hooshmand, K. Adeli, Principal component and correlation analysis of biochemical and endocrine markers in a healthy pediatric population (CALIPER), Clin. Biochem. 66 (2019) 29–36, https://doi.org/10.1016/j.clinbiochem.2019.02.004.
[18] E. Grossi, R. Colombo, S. Cavuto, C. Franzini, The REALAB project: a new method for the formulation of reference intervals based on current data, Clin. Chem. 51 (2005) 1232–1240, https://doi.org/10.1373/clinchem.2005.047787.
[19] T. Kouri, V. Kairisto, A. Virtanen, E. Uusipaikka, A. Rajamäki, H. Finneman, et al., Reference intervals developed from data for hospitalized patients: computerized method based on combination of laboratory and diagnostic data, Clin. Chem. 40 (1994) 2209–2215, https://doi.org/10.1093/clinchem/40.12.2209.
[20] A.K. Aarsand, J. Díaz-Garzón, P. Fernandez-Calle, E. Guerra, M. Locatelli, W. A. Bartlett, et al., The EuBIVAS: within- and between-subject biological variation data for electrolytes, lipids, urea, uric acid, total protein, total bilirubin, direct bilirubin, and glucose, Clin. Chem. 64 (2018) 1380–1393, https://doi.org/10.1373/clinchem.2018.288415.
[21] ICMJE, Recommendations. https://icmje.org/recommendations/, 2024 (accessed 18 April 2024).
[22] E. von Elm, D.G. Altman, M. Egger, S.J. Pocock, P.C. Gøtzsche, J. P. Vandenbroucke, Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies, Lancet. 370 (2007) 1453–1457, https://doi.org/10.1016/S0140-6736(07)61602-X.
[23] S. Heidari, T.F. Babor, P. De Castro, S. Tort, M. Curno, Sex and Gender Equity in Research: rationale for the SAGER guidelines and recommended use, Res. Integr. Peer Rev. 1 (2016) 2, https://doi.org/10.1186/s41073-016-0007-6.
[24] S.M. Schüssler-Fiorenza Rose, K. Contrepois, K.J. Moneghetti, W. Zhou, T. Mishra, S. Mataraso, et al., A longitudinal big data approach for precision health, Nat. Med. 25 (2019) 792–804, https://doi.org/10.1038/s41591-019-0414-6.
[25] J.G. Cohen, A.Q. Tran, B.J. Rimel, I. Cass, C.S. Walsh, B.Y. Karlan, et al., Thrombocytosis at secondary cytoreduction for recurrent ovarian cancer predicts suboptimal resection and poor survival, Gynecol. Oncol. 132 (2014) 556–559, https://doi.org/10.1016/j.ygyno.2014.01.003.
[26] A. Digklia, I.A. Voutsadakis, Thrombocytosis as a prognostic marker in stage III and IV serous ovarian cancer, Obstet. Gynecol. Sci. 57 (2014) 457–463, https://doi.org/10.5468/ogs.2014.57.6.457.
[27] H. Eggemann, J. Ehricke, T. Ignatov, F. Fettke, A. Semczuk, S.D. Costa, et al., Platelet count after chemotherapy is a predictor for outcome for ovarian cancer patients, Cancer Invest. 33 (2015) 193–196, https://doi.org/10.3109/07357907.2015.1020384.
[28] K. Adeli, J.E. Raizman, Y. Chen, V. Higgins, M. Nieuwesteeg, M. Abdelhaleem, et al., Complex biological profile of hematologic markers across pediatric, adult, and geriatric ages: establishment of robust pediatric and adult reference intervals on the basis of the Canadian health measures survey, Clin. Chem. 61 (2015) 1075–1086, https://doi.org/10.1373/clinchem.2015.240531.
[29] G. Biino, I. Santimone, C. Minelli, R. Sorice, B. Frongia, M. Traglia, et al., Age- and sex-related variations in platelet count in Italy: a proposal of reference ranges based on 40987 subjects' DATA, PLOS ONE. 8 (2013) e54289.
[30] R. Qayyum, B.M. Snively, E. Ziv, M.A. Nalls, Y. Liu, W. Tang, et al., A meta-analysis and genome-wide association study of platelet count and mean platelet volume in African Americans, PLOS Genet. 8 (2012) e1002491.
[31] M.B. Kucukay, R. Alanli, Vitamin D replacement effect on platelet counts, J. Coll. Phys. Surg. Pak. 31 (2021) 1064–8. 10.29271/jcpsp.2021.09.1064.
[32] M. Schlesinger, Role of platelets and platelet receptors in cancer metastasis, J. Hematol. Oncol. 11 (2018) 125, https://doi.org/10.1186/s13045-018-0669-2.
[33] D.G. Menter, S.C. Tucker, S. Kopetz, A.K. Sood, J.D. Crissman, K.V. Honn, Platelets and cancer: a casual or causal relationship: revisited, Cancer Metastas. Rev. 33 (2014) 231–269, https://doi.org/10.1007/s10555-014-9498-0.
[34] M. Kim, H. Chang, H.C. Yang, Y.J. Kim, C.T. Lee, J.H. Lee, et al., Preoperative thrombocytosis is a significant unfavorable prognostic factor for patients with resectable non-small cell lung cancer, World. J. Surg. Oncol. 12 (2014) 37, https://doi.org/10.1186/1477-7819-12-37.
[35] A.S. Chadha, E. Kocak-Uzel, P. Das, B.D. Minsky, M.E. Delclos, U. Mahmood, et al., Paraneoplastic thrombocytosis independently predicts poor prognosis in patients with locally advanced pancreatic cancer, Acta Oncol. 54 (2015) 971–978, https://doi.org/10.3109/0284186X.2014.1000466.
[36] C. Hu, R. Chen, W. Chen, W. Pang, X. Xue, G. Zhu, et al., Thrombocytosis is a significant indictor of hypercoagulability, prognosis and recurrence in gastric cancer, Exp. Ther. Med. 8 (2014) 125–132, https://doi.org/10.3892/etm.2014.1699.
[37] P. Adibi, E. Faghih Imani, M. Talaei, M. Ghanei, Population-based platelet reference values for an Iranian population, Int. J. Lab Hematol. 29 (2007) 195–199, https://doi.org/10.1111/j.1751-553X.2006.00843.x.
[38] D. Menard, M.J. Mandeng, M.B. Tothy, E.K. Kelembho, G. Gresenguet, A. Talarmin, Immunohematological reference ranges for adults from the Central African Republic, Clin. Vaccin Immunol. 10 (2003) 443–445, https://doi.org/10.1128/CDLI.10.3.443-445.2003.
[39] L. Peng, J. Yang, X. Lu, T. Okada, T. Kondo, C. Ruan, et al., Effects of biological variations on platelet count in healthy subjects in China, Thromb. Haemost. 91 (2004) 367–372, https://doi.org/10.1160/TH03-05-0276.
[40] J. Rolland, D. Eveillard, B. Delahaye, R. Boutin, Datascape: Exploring heterogeneous dataspace, Sci. Rep. 14 (2024) 7041, https://doi.org/10.1038/s41598-024-52493-7.