Journal of Biomedical Informatics 131 (2022) 104111

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Individual reference intervals for personalised interpretation of clinical and metabolomics measurements

Murih Pusparum [a,b,*], Gökhan Ertaylan [b,*], Olivier Thas [a,c,d,*]

[a] Data Science Institute, I-Biostat, Hasselt University, Hasselt 3500, Belgium
[b] Health, Flemish Institute for Technological Research (VITO), Mol 2400, Belgium
[c] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent 9000, Belgium
[d] National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, Wollongong 2500, NSW, Australia

ARTICLE INFO

Keywords:
Reference intervals
Linear quantile mixed models
Penalised estimation
Precision health
Personalised medicine
Biological variation

ABSTRACT

The Population Reference Interval (PRI) refers to the range of outcomes that are expected in a healthy population for a clinical or a diagnostic measurement. It is widely used in daily clinical practice and is essential for assisting clinical decision-making in diagnostics and treatment. In this manuscript, we start from the observation that each healthy individual has its own range for a given variable, depending on personal biological traits. This Individual Reference Interval (IRI) can be calculated and be utilised in clinical practice, in combination with the PRI for improved decision making. Nonparametric estimation of IRIs would require quite long time series. To circumvent this problem, we propose methods based on quantile models in combination with penalised parameter estimation methods that allow for information-sharing among the subjects. Our approach considers the calculation of an IRI as a prediction problem rather than an estimation problem. We perform a simulation study designed to benchmark the methods under different assumptions. From the simulation study we conclude that the new methods are robust and provide empirical coverages close to the nominal level. Finally, we evaluate the methods on real-life data consisting of eleven clinical tests and metabolomics measurements from the VITO IAM Frontier study.

1. Introduction

In the early days of clinical practice diagnostic, (laboratory) tests were developed to be performed with basic equipment, resulting in rapid evaluation of the results, and the rendering of a diagnostic opinion. Choice of an appropriate test, its performance, and interpretation entirely depended on the practitioner [1].

In today's medical practice, clinical laboratory tests are routinely performed by certified clinical laboratories for examining the clinical, physiological or molecular state of a patient. Reference intervals (RI) are essential for the interpretation of clinical laboratory tests, assisting the professionals regarding diagnosis and decision making in patient care. Such intervals are calculated from a reference population of healthy individuals and will be referred as *Population Reference Intervals* (PRI) [2]. For any particular clinical marker, the specific PRI is compared with the laboratory test results from the patient; when the result is within the PRI a normal reading is declared. If the test results are outside the PRI boundaries, the context of the patient explaining the results and other tests are taken into account to determine a course of action. The practitioners would also often look at the difference between the test results and the historical measurements; whether there is an unexpected change or a monotonic trend that may indicate abnormalities, even though the test results are still within the PRI. This type of trend assessment of the data should be done prior to the RI estimation but not within the area this study.

Besides the PRI, clinical decision limits (CDLs) are also often

Abbreviations: RI, reference interval; PRI, population reference interval; IRI, individual reference interval; CDL, clinical decision limit; IFCC, International Federation of Clinical Chemistry and Laboratory Medicine; LQMM, linear quantile mixed model; JQM, joint quantile model; PJQM2, penalised joint quantile model with two subject-specific effects; PJQM1, penalised joint quantile model with one subject-specific effect; TEC, time empirical coverage; SEC, subject empirical coverage; LMM, linear mixed model; PCA, principal component analysis.

* Corresponding authors at: Data Science Institute, Hasselt University, Hasselt 3500, Belgium (M. Pusparum, O. Thas), Health Unit, Flemish Institute for Technological Research (VITO), Mol 2400, Belgium (G. Ertaylan).

E-mail addresses: murih.pusparum@uhasselt.be (M. Pusparum), gokhan.ertaylan@vito.be (G. Ertaylan), olivier.thas@uhasselt.be (O. Thas).

https://doi.org/10.1016/j.jbi.2022.104111
Received 15 December 2021; Received in revised form 22 April 2022; Accepted 1 June 2022
Available online 4 June 2022

reported for some clinical tests. While the PRIs describe the distribution in an apparently healthy population and hence can assist in separating the healthy individuals from the unhealthy ones, the CDLs can define and associate a significantly higher risk of adverse outcome or an onset of a specific disease [3]. The CDL is composed by one clinical threshold, usually an upper limit, computed based on clinical outcome studies (prospective cohort studies, meta-analysis). The typical clinical tests that report the CDLs values include some lipid parameter: total cholesterol, low-density lipoprotein cholesterol (LDL cholesterol), high-density lipoprotein cholesterol (HDL cholesterol), among others. Other clinical parameters with CDLs relate to diabetes intervention (hemoglobin A1c (HbA1c), fasting glucose) [3]. Although clear differences exist between the PRIs and CDLs, and the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) also recommends to only report CDLs (when available), Following the reviewer's sugestion the current clinical practice for most tests involves reporting the PRIs for the clinical interpretation.

At present, the PRI estimation methods [4–6] typically require cross-sectional data, i.e. data that consist of one measurement for each subject in the sample. The latter is assumed to represent a population of healthy people. In general, the methods can be grouped into the parametric and the non-parametric methods. The former relies on distributional assumptions, while the latter simply estimates the PRI bounds by the appropriate sample order statistics (i.e. the empirical quantiles). In practice, many PRIs are computed using the non-parametric methods as they are free of distributional assumptions, but such methods require sufficiently large samples. An example of this is the method to obtain PRIs by using the existing clinical and laboratory data from the electronic medical record, and by applying a *posteriori* approach it can extract the reference intervals associated with ICD9 codes diseases [7]. Combined with the bootstrap resampling technique, the non-parametric methods may provide a robust estimate and it fulfills the recommendations of IFCC [8,9]. A study has been presented to compare these classical methods for PRI estimations on a large cross-sectional data, which showed that the non-parametric methods are more preferable [2]. These methods, however, are not appropriate when data comes as time-series as they do not adequately capture the within and between-subjects variability.

Despite the widespread use of the PRI in daily clinical practice, it falls short in providing the individual context necessary to recognize diseases at an early stage. Precision health starts from the premise that each person has its individual specification of "healthy", resulting from individual biological traits. This hypothesis is also supported by the real-life data of healthy individuals, that each individual's clinical test results and metabolomics measurements tend to cluster together, separated from the peers, such as shown in Fig. 1. From this perspective,

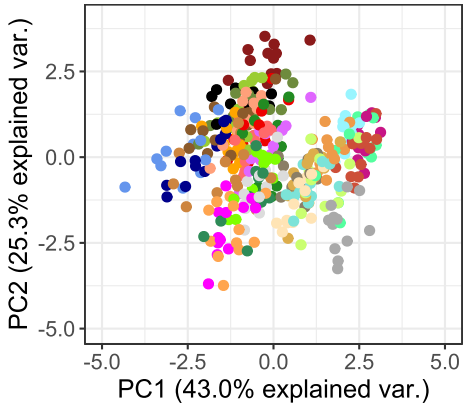

Fig. 1. The first two principal components of 11 clinical tests and metabolomics measurements in the IAM Frontier dataset. Different colors represent different individuals. Data from one individual are clustered together.

definition of personalised health would benefit from PRIs that are subject-specific. We refer to such intervals as *Individual Reference Intervals* (IRI). Such an interval would indicate what test results can be expected with e.g. a probability of 95%, when that person is in its healthy state. Since the number of available data from a single subject is typically small, the conventional methods for PRI calculation cannot be used for the IRI. Recently, Coşkun et al. [10] proposed a method that allows for the calculation of IRIs starting from time series data. For a 100α% IRI for subject i, their lower and upper bounds are computed as $\overline{y}_i \pm z_{0.5+\alpha/2}\mathrm{TV}\overline{y}_i$, where $\overline{y}_i$ is the sample mean of the time series data of subject i, $z_{0.5+\alpha/2}$ is the $100(0.5+\alpha/2)$ percentile of a standard normal distribution (α is the nominal coverage probability $\tau_2 - \tau_1$) and TV is the square root of the total variation which they express as the sum of the squared coefficients of variation of the biological variation and the analytical imprecision. However, from this description, particularly from its symmetric construction and the use of the standard normal distribution, it is clear that it is a parametric method that relies on the normal distribution of the clinical test results. In this paper, we will propose nonparametric methods for the calculation of IRIs that do not rely on distributional assumptions, starting from rather short time-series from multiple subjects. We will approach the problem as a prediction problem.

Three procedures will be described for constructing the IRIs, allowing for both variation within and between subjects, and requiring time-series data of several subjects for model fitting. The first approach makes directly use of Linear Quantile Mixed Models (LQMM) and has been proposed recently [2], but it has not been properly evaluated yet. This method only allows for the separate estimation of the lower and upper bounds of the IRIs. Our second method is based on a new Joint Quantile Model (JQM) that simultaneously models the subject-specific lower and upper bounds. A penalised parameter estimation procedure is proposed, which allows for considering the subject-specific effects to be fixed and for the calculation of IRIs. For the calibration of the penalty parameters we take the perspective of predictive modelling. The last method is based on a quantile model that was originally proposed for longitudinal data [11] and for which the subject-specific effects are also fixed and estimated by a penalised estimation method. The optimisation of the penalty parameter makes use of the same procedure that we developed for the JQM method. A major difference with the JQM method is that it does not produce IRIs with subject-specific lengths.

In Section 2, we describe the materials and the methods with corresponding model formulations. In Section 3, the methods are evaluated, both using simulated data as well as the real-life cohort study data. This cohort study data has been collected by The Flemish Institute for Technological Research (VITO) in a pilot time-series study in which, during 12 months, 30 healthy individuals donated blood, urine, and stool samples at monthly visits [12]. These samples were collected after an overnight fasting for at least eight hours. At bimonthly visits, both their metabolomics measurements and clinical tests were taken and assessed by accredited labs and appointed doctors. Finally, a conclusion is formulated in Section 4.

Throughout the paper, we will use different terms such as: 1) *clinical tests/clinical test results*, the results of the medical test performed from blood and urine samples of participants, 2) *metabolomics measurements*, the results of the targeted NMR metabolomics performed using the Nightingale platform for the analysis of blood plasma, and 3) *model parameters*, any values that describe an aspect of a population and need to be estimated from the statistical models.

2. Materials and methods

2.1. VITO IAM frontier data

The VITO IAM Frontier study contains several types of data, including clinical, omics (proteomics, metabo-lomics, genomics), and

physical characteristics of 30 healthy individuals that were measured throughout a one year period. In this paper we focus on eleven clinical tests and metabolomics measurements, assessed by two independent accredited laboratories (see Section 4 in the *Supplementary Document* for descriptions). These include glucose, triglyceride, total cholesterol, LDL, HDL, and non-HDL cholesterol, albumin, creatinine, apolipoprotein A1, apolipoprotein B, and the ratio of apolipoprotein B to apolipoprotein A1. The similarity of clinical tests and corresponding metabolomics measurements is very high since in principle they measure the same chemical compound presents in the body. Although a reasonable difference was still observed as different technologies were used, but most of both results have the same interpretation w.r.t. the PRIs. Table 1 gives a general overview of the data and a similarity matrix are presented in Figure S15 in the *Supplementary Document*.

2.2. Principal component analysis

Initial data exploration was done to the IAM Frontier dataset using the Principal Component Analysis (PCA). Fig. 1 shows the first and the second principal components (PCs) implemented in the standardised values of eleven clinical tests and metabolomics measurements. These two PCs already explain 68% of the variance, where data from the same individual are clearly clustered together. This indicates that the healthy definition can differ from one individual to another, demonstrating the importance of calculating the IRI.

2.3. Linear quantile mixed model

2.3.1. Model description

In this section we describe a method based on separately fitting two LQMMs [14,15] that only include a fixed intercept and a subject-specific random intercept to model the between-subject variability of the reference intervals. This is essentially not a new method; it has been proposed earlier [2] for IRI calculation, but it has not been properly evaluated yet.

Consider the quantile function $Q(\tau)$ of a random variable Y. This is defined for all $\tau \in [0,1]$ as $Q(\tau) = \inf\{y \in \mathbb{R} : F(y) \geqslant \tau\}$. The index i is added to the quantile function, i.e. $Q_i(\tau)$, to make it refer to the distribution of the measurement Y_i of subject $i = 1,\ldots,N$ in the dataset. The probabilities τ_1 and $\tau_2 > \tau_1$ are chosen so as to make $[Q_i(\tau_1), Q_i(\tau_2)]$ the IRI of subject i with nominal coverage probability $\tau_2 - \tau_1$. The quantile model for subject i then becomes

$$Q_i(\tau_1) = \beta_{01} + u_{1i} \tag{1}$$

$$Q_i(\tau_2) = \beta_{02} + u_{2i}, \tag{2}$$

where $\beta_{01}, \beta_{02} \in \mathbb{R}$ are the fixed intercepts, and u_{1i} and u_{2i} are the subject-specific random effects. In this paper we assume these random intercepts u_{1i} and u_{2i} to follow zero-mean normal distributions with variances ψ_{u1} and ψ_{u2}, respectively. Moreover, the random intercepts are assumed to be independently distributed. For the estimation of the parameters, we follow the procedure of quantile regression models with only a random intercept [14,15] and use the implementation in the *lqmm* R package [16].

2.3.2. IRI for a new subject

Suppose we have estimates of all β parameters and predictions of all u_is (u_{1i} and u_{2i}) from all subjects i ($i = 1,\ldots,N$). We now aim to compute the IRI of a new subject, say subject m for which some data is already available, but not used as part of the training data for fitting the quantile model. We can still use the estimates $\widehat{\beta}_{01}$ and $\widehat{\beta}_{02}$ as fixed effects, but we need predictions of the new subject-specific effects u_{1m} and u_{2m}. These predictions are obtained as the mean of the posterior distribution $[u_m|\mathbf{y_m}, \widehat{\beta}]$, where $\mathbf{y_m}$ denotes the available historical data from the new subject m and $\widehat{\beta}$ denotes the estimated intercepts from model (1) and (2). This procedure is a modification of the estimated best linear predictor (eBLP) of the LQMM [16] and is explained in detail in Section 1 in the *Supplementary Document*. We adapted this procedure so that we only use $\mathbf{y_m}$ instead of the data used for fitting the quantile model Y_i.

2.4. Joint quantile model

2.4.1. Model description

In this section we introduce a Joint Quantile Model (JQM) which forms the basis for simultaneously estimating the bounds of the IRIs. Upon using the same notation as before, we define the JQM by the following three equations ($i = 1,\ldots,N; j = 1,\ldots,n_i$):

$$Q_i(0.5) = \beta_0 + u_i \tag{3}$$

$$Q_i(\tau_1) = \beta_0 + u_i + z_i\beta_1 \tag{4}$$

$$Q_i(\tau_2) = \beta_0 + u_i + z_i\beta_2 \tag{5}$$

where β_0 and the subject-specific effects u_i and z_i are shared among all models (hence the term *joint* quantile model).

According to the model the median of the data of subject i is given by $\beta_0 + u_i$. Similar as with the LQMM in Section 2.3.1, with τ_1 and $\tau_2 > \tau_1$ such that $\tau_1 + \tau_2 = 1$, we can interpret the quantiles in (4) and (5) as the lower and the upper bounds of an IRI for subject i, with coverage probability $\tau_2 - \tau_1$. The length of the IRI is given by $Q_i(\tau_2) - Q_i(\tau_1) =$

Table 1
Subject demographics and data description of 11 metabolomics measurements and clinical tests in the IAM Frontier study. The medians are presented together with the interquartile ranges (IQR) in parentheses.

Characteristic (Median (IQR))	N = 30 individuals, n = 6 time points		
	Clinical	Metabolomics	PRI (Lower, Upper bound) [13]
Gender	15 males, 15 females		
Age (years)	51 (47, 54)		
Albumin (g/l)	46.00 (44.00, 48.00)	41.91 (40.12, 43.6)	(35, 50)
Apolipoprotein A1 (g/l)	1.70 (1.48, 1.91)	1.55 (1.34, 1.70)	(0.80, 1.70)
Apolipoprotein B (g/l)	0.96 (0.79, 1.08)	0.90 (0.77, 1.01)	(0.25, 1.20)
Apolipoprotein B/A1 (ratio)	0.57 (0.45, 0.70)	0.60 (0.46, 0.72)	
Creatinine (μmol/l)	78.23 (70.72, 91.05)	81.53 (72.61, 94.1)	(53, 106)
Glucose (mmol/l)	5.11 (4.77, 5.49)	5.19 (4.89, 5.64)	(3.90, 6.10)
LDL cholesterol (mmol/l)	3.20 (2.69, 3.68)	2.88 (2.34, 3.30)	(0, 4.12)
HDL cholesterol (mmol/l)	1.61 (1.22, 1.84)	1.51 (1.15, 1.73)	(1.03, 1.55)
non-HDL cholesterol (mmol/l)	3.81 (3.21, 4.27)	3.55 (3.06, 4.04)	(0, 4.00)
Total cholesterol (mmol/l)	5.34 (4.77, 5.86)	5 (4.50, 5.48)	(0, 6.18)
Triglycerides (mmol/l)	1.12 (0.78, 1.53)	1.20 (0.83, 1.61)	(0.11, 2.15)

*IQR, interquartile range; PRI, population reference interval; LDL, low-density lipoprotein; HDL, high-density lipoprotein.

$z_i(\beta_2 - \beta_1)$. Hence, this model does not only allow for subject-specific locations of the reference intervals (given by u_i), but it also allows for subject-specific lengths. In this respect, the z_is are individual scaling factors restricted to be non-negative ($z_i \geqslant 0$), and $\beta_0 + \beta_1$ and $\beta_0 + \beta_2$ are the lower and upper bounds of an IRI of a subject with $u_i = 0$ and $z_i = 1$.

Since parameters are shared in three quantiles, these parameters will need to be estimated by considering the three models simultaneously. Several approaches for estimating the model parameters can be considered. For example, we could have assumed that the subject-specific u_is and z_is are random effects, distributed according to a user-specified distribution (e.g. normal distribution for u_i and inverse gamma for z_i). From this perspective, the u_is act as random intercepts such as in the LQMM. However, the terms $z_i\beta_1$ and $z_i\beta_2$ with z_i as a random effect, do not fit into the class of LQMMs. Moreover, in the LQMM approach the distribution of the random effects would need to be specified and therefore the method would enforce a distributional assumption which we want to avoid. While the LQMM literature often focuses on the inference of the model parameters, our objective is to construct IRIs that have the correct probabilistic interpretation when applied to new test results of subjects. From this perspective, our objective is closer related to predictive modelling than to inference. For these reasons, we do not follow the LQMM approach. See Section 3 of the *Supplementary Document*. Instead we consider the subject-specific u_is and z_is as fixed parameters and hence they will have to be estimated from the training data. With this large number of parameters, and knowing that the training data only contains short time series, we propose a penalised estimation procedure with ℓ_2 penalty, in which the penalty parameters are calibrated to make the coverage of the IRIs close to the nominal coverage level $\tau_2 - \tau_1$. The details of this procedure are deferred to the next section. We refer to this JQM method with types of two subject-specific effects (u_i and z_i) estimated by a penalised procedure as the Penalised JQM2 (PJQM2).

2.4.2. Parameter estimation

We propose to estimate β_0 as the median of all data. Further, we estimate $\beta_k, k = 1, 2$ and the subject-specific effects by minimising the function

$$M(\beta, \mathbf{u}, \mathbf{z}; \lambda) = \sum_{k=1}^{2} \sum_{i=1}^{N} \sum_{j=1}^{n_i} \rho_{\tau_k}(y_{ij} - \beta_0 - u_i - z_i\beta_k) + \lambda_u \sum_{i=1}^{N} u_i^2 + \lambda_z \sum_{i=1}^{N} (z_i - 1)^2 \tag{6}$$

in which $\rho_\tau(w) = w(\tau - w \leqslant 0)$ is the check function. The first term corresponds to the objective function for estimating parameters in linear quantile models, and the two last terms are ℓ_2 penalty terms with user-specified penalty parameters $\lambda_u, \lambda_z \geqslant 0$. The first penalty term forces the subject-specific effects u_i to zero as λ_u increases (i.e. the subject-specific medians are shrunk towards one another), and the scaling factors z_i are shrunk towards 1 as λ_z increases, bringing the lengths of the IRIs closer to one another.

The function M in (6) can also be considered as a log-likelihood function of a LQMM. The penalty terms then arise if the u_i and z_i were considered as normal random effects with mean zero and one, respectively, and the penalty parameters then correspond to the (inverse) variance parameters. However, the theory of LQMMs do not accommodate for terms of the form $z_i\beta_k$ [14,15], and our focus is on prediction (in the sense of Section 3 of the *Supplementary Document*). Our motivation is twofold. First, our model specification requires restrictions on the parameters to make them identifiable. This is accomplished by the penalisation of the u_is and z_is. Second, the penalisation terms (or shrinkage) cause information sharing between subjects. This is necessary, because in realistic settings the numbers of replicates per subject (n_i) are typically quite small, too small to nonparametrically estimate e.g. the 2.5% and 97.5% quantiles of IRIs. The user-specified penalty parameters will be selected by means of a cross-validation procedure aiming at calibrating the IRIs at the nominal coverage probability $\tau_2 - \tau_1$ (see further down).

In order to minimise the objective function for a given λ_u and λ_z, and obtain the estimates of all model parameters, we propose an iterative procedure. This iterative procedure is explained more in detail in Section 2 in the *Supplementary Document*. The procedure for selecting the penalty parameters is described in Section 2.4.4. First we need to show how IRIs are computed for new subjects, i.e. for a subject that did not contribute its data for the estimation of the parameters, but for which some historical data is available.

2.4.3. IRI for a new subject

Suppose we have estimates of all β parameters and of all u_is and z_is and that the penalty parameters λ_u and λ_v have been optimised as will be described in the next section. We now aim to compute the IRI of a new subject, say subject m. We can still use the estimates $\widehat{\beta}_0, \widehat{\beta}_1$ and $\widehat{\beta}_2$, but we need estimates of the new subject-specific effects u_m and z_m. Our method proceeds along the following steps:

1. Set $\widehat{z}_m = 1$ (as if the new subject is an *average* subject).
2. Estimate u_m by minimising

$$\sum_{k=1}^{2} \sum_{j=1}^{n_m} \rho_{\tau_k}(y_{mj} - \widehat{\beta}_0 - u_m - \widehat{z}_i\widehat{\beta}_k) + \lambda_u u_m^2.$$

   The solution is denoted by $\widehat{u}_m$
3. Estimate z_m by minimising

$$\sum_{k=1}^{2} \sum_{j=1}^{n_m} \rho_{\tau_k}(y_{mj} - \widehat{\beta}_0 - \widehat{u}_m - z_m\widehat{\beta}_k) + \lambda_z (z_m - 1)^2.$$

   The solution is denoted by $\widehat{z}_m$.
4. Iterate over the two previous steps until convergence.

The IRI of subject m is then given by

$$[\widehat{\beta}_0 + \widehat{u}_m + \widehat{z}_m\widehat{\beta}_1, \widehat{\beta}_0 + \widehat{u}_m + \widehat{z}_m\widehat{\beta}_2].$$

Note that the IRI of a new subject can only be computed if the subject comes with at least a few data points.

2.4.4. Selection of the penalty parameters

The selection of penalty parameters proceeds along the lines of a leave-one-out cross validation scheme. We use two types of coverage probabilities (Time Empirical Coverage (TEC) and the Subject Empirical Coverage (SEC)) for assessing the performance of the algorithm and hence select the penalty parameters that result in the optimal coverage of intervals. In our context, an empirical coverage is an approximation of the probability that an interval covers an individual clinical test result. A good method gives a coverage close to the nominal level (e.g. 95%). In particular, the TEC corresponds to the coverage of a new measurement of subjects that have their historical data in the training data, and the SEC corresponds to the coverage of a new measurement of a new subject that did not contribute to the training data. These training data are used for estimating all parameters in the models. These two types of coverages are highly relevant for our purposes, as they are directly related to the definition of reference intervals in terms of a nominal coverage, and to the practical use of such intervals. We explain this algorithm in detail in the following steps. Consider two sets of possible values for λ_u and λ_z, say L_u and L_z. For each $\lambda_u \in L_u$ and each $\lambda_z \in L_z$, the steps are performed for each subject $i = 1, \ldots, N$.

1. Compute the Time Empirical Coverage (TEC) by splitting the dataset into:

- Dataset D_l: dataset with all data from the last clinical test result from all subjects.
- All remaining data. This dataset is referred to as the *training dataset*, illustrated as the bold font face in the matrix below.

$$\begin{pmatrix} \mathbf{y_{11}} & \mathbf{y_{21}} & \cdots & \mathbf{y_{(N-1)1}} & \mathbf{y_{N1}} \\ \mathbf{y_{12}} & \mathbf{y_{22}} & \cdots & \mathbf{y_{(N-1)2}} & \mathbf{y_{N2}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{y_{1(n-1)}} & \mathbf{y_{2(n-1)}} & \cdots & \mathbf{y_{(N-1)(n-1)}} & \mathbf{y_{N(n-1)}} \\ y_{1n} & y_{2n} & \cdots & y_{(N-1)n} & y_{Nn} \end{pmatrix}$$

Use the training data for estimating all parameters and compute the empirical coverage based on the observations in D_l. This is calculated as the relative frequency of observations in D_l that are contained in the IRIs of the corresponding subjects.

2. For a given subject i, split the dataset into four datasets:
   - Dataset D_i^1: dataset with all data from subject i, except for the last observation.
   - Dataset D_{li}^1: The last observation of D_i^1.
   - Dataset D_i^2: dataset with all data from the last clinical test results of all subjects except for subject i, i.e. the observations $y_{ln_l}, l \neq i$.
   - All remaining data. This dataset is referred to as the *training dataset*, illustrated as the bold font face in the matrix below, for $i = N$. In this matrix, $D_{li}^1 = y_{Nn}$.

$$\begin{pmatrix} \mathbf{y_{11}} & \mathbf{y_{21}} & \cdots & \mathbf{y_{(N-1)1}} & y_{N1} \\ \mathbf{y_{12}} & \mathbf{y_{22}} & \cdots & \mathbf{y_{(N-1)2}} & y_{N2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{y_{1(n-1)}} & \mathbf{y_{2(n-1)}} & \cdots & \mathbf{y_{(N-1)(n-1)}} & y_{N(n-1)} \\ y_{1n} & y_{2n} & \cdots & y_{(N-1)n} & y_{Nn} \end{pmatrix}$$

Use the training data for estimating all parameters. Compute the empirical coverage based on D_{li}^1, given subject i's historical data D_i^1. This is a binary indicator of whether D_{li}^1 is contained in the IRI for subject i (since subject i was not part of the training data, its IRI is calculated as outlined in Section 2.4.3. The relative frequency of this indicator over all subjects will be referred to as the Subject Empirical Coverage (SEC).

Upon completion of steps 1 and 2 for all subjects $i = 1, \ldots, N$, the optimal (λ_u, λ_z) is then selected as the parameter combination resulting in the smallest of the empirical coverages TEC and SEC (i.e. min(TEC, SEC)), that is at least as large as the nominal coverage level.

2.5. Joint quantile model for IRIs with constant length

For the analysis of longitudinal data, Koenker [11] proposed a joint quantile model that also includes subject-specific effects as fixed effects. The parameters are also estimated by means of a penalised estimation procedure, but using an ℓ_1 penalty term instead of the ℓ_2 penalty that we introduced in the previous section. In the original paper [11], the penalty parameter was assumed fixed, but here we adopt our cross-validation method that we introduced for the PJQM2 method.

The penalised Joint Quantile Model (JQM) of Koenker [11] also simultaneously models the IRI's boundaries. We define the following models

$$Q_i(\tau_1) = \beta_{01} + u_i \tag{7}$$

$$Q_i(\tau_2) = \beta_{02} + u_i, \tag{8}$$

where $\beta_{01}, \beta_{02} \in \mathbb{R}$ are the fixed intercepts, and u_i is the shared subject-specific effect. The same as before, the quantile model in (7) and (8) act as the lower and the upper bounds of an IRI for subject i. The length of the IRI is further given by $Q_i(\tau_2) - Q_i(\tau_1) = \beta_{02} - \beta_{01}$. We now see that the model only gives subject-specific locations of IRIs, but the length is constant for all subjects. We further refer this method as the Penalised JQM with one subject-specific effect (PJQM1).

The parameter estimates are obtained by minimising the following objective function [11]

$$M(\beta, \mathbf{u}; \lambda) = \sum_{k=1}^{2} \sum_{i=1}^{N} \sum_{j=1}^{n_i} w_k \rho_{\tau_k}(y_{ij} - \beta_{0\tau_k} - u_i) + \lambda \sum_{i=1}^{N} |u_i|,$$

where $\lambda \geqslant 0$ is the penalty parameter. An equal weight $w_k = 0.5$ is assigned so that $\sum_{k=1}^{2} w_k = 1$. We use the implementation in the *rqpd* R package [17] to obtain the estimates of $\beta_{0\tau_k}$ and u_i.For the selection of the penalty parameter λ and the IRI estimation of a new subject, we follow the same procedure as described in Sections 2.4.4 and 2.4.3, respectively.

The summary of the proposed methods for constructing the IRIs is presented in Table 2.

3. Results

3.1. Simulated data

We conducted a simulation study to evaluate the performance of IRI computed using LQMM, PJQM2, and PJQM1 methods. Data were generated with linear mixed models (LMM),

$$y_{ij} = \beta_0 + u_i + \varepsilon_{ij} \quad i = 1, \ldots, N, j = 1, \ldots n,$$

with β_0 fixed at 0 and the u_i are random intercepts. We considered both a standard normal distribution and a scaled χ_4^2 distribution for simulating the u_i. The latter is scaled so that is has mean 0 and variance 1. Since both the LMM and the quantile models are location-shift models, β_0 and the u_is from the LMM correspond to the β_0 and the u_is of our quantile models. For the error terms we considered a normal distribution with mean zero and variance θ_i^2. This variance was either set to a constant ($\theta_i^2 = 1$), or it was sampled from an inverse gamma distribution with shape parameter a and scale parameter b. The latter generates data with subject-specific variances (heteroscedasticity) and hence subject-specific lengths of the IRI. Table S1 in the *Supplementary Document* gives a description of all simulation scenarios. The choice of a and b values were derived from several VITO IAM Frontier clinical tests and hence represents realistic settings.

For each scenario, 100 Monte Carlo simulation runs were performed. For each simulation run, data for $N+1$ subjects with $n+1$ repeated measurements were simulated. The data for the first N subjects and their first n repeated measurements were used for estimating the model parameters. The data for the $N+1$th subject and for the $(n+1)$th repeated measurement were used for the estimation of the empirical coverages. The generation of data for an additional subject and for an additional time point corresponds to two possible applications in real life settings: 1) when existing subjects have new measurements, and 2) when there is a new subject that was not included in the data used for the parameter

Table 2
Summary of the models and methods described in this paper.

Method	Model	Estimation procedure	Subject-specific effects	IRI length
LQMM	2 separate LQMMs	least abs. dev. (LAD)	u_i random	varying
PJQM2	Joint quantile model	LAD + ℓ_2-penalisation	u_i and z_i fixed	varying
PJQM1	Joint quantile model	LAD + ℓ_1-penalisation	u_i fixed	constant

estimation. As before, we refer to the empirical coverage that corresponds to the first case as the TEC, and to the second as the SEC.

No competitor methods are included in this evaluation study, because to our knowledge there are no nonparametric competitors that can work with a longitudinal data structure with rather short time series. Although the method of Coşkun et al. [10] is applicable to this data structure, we did not include it here as it is a parametric method relying on normality. The interested reader, however, may find some simulation results in Section 5 in the *Supplementary Document*. From this simulation, we found that this method fails detecting an unexpected small test results in a data with a highly-skewed distribution: when there are small new measurements, the lower bound would always detect them as within the interval.

Fig. 2 shows the TEC and SEC for all scenarios at 95% nominal coverage. A method with a good performance would have TEC and SEC approximately equal to this nominal coverage. For all numbers of repeated measurements n and numbers of subjects N, the TECs of the IRI computed by PJQM1 are consistently close to the nominal coverage level. The TECs of LQMM and PJQM2, on the other hand, are slightly smaller than the nominal coverage, but they increase as n becomes larger. We do not observe a strong effect of n on the SEC, although the SECs of PJQM2 and PJQM1 are often higher and closer to the nominal coverage level than LQMM. We also observe that the number of subjects and random effect distributions do not strongly affect the performance. The other TEC and SEC values at different nominal coverages can be consulted in Figures S1 and S2 in the *Supplementary Document*.

The comparison of IRI lengths estimated by the three methods is shown in Fig. 3 for the scenario with a standard normal random effects $N(0,1)$ and normal error terms $N(0,\theta_i)$ where θ_i was sampled from an Inverse-Gamma(2,2.5). While PJQM2 gives varying lengths for different subjects, the PJQM1 only have one fixed length. From 20 simulated datasets, we observe that PJQM1 gives larger lengths than LQMM and PJQM2, and hence wider intervals (*left and middle panels*). However, when we compare them to LQMM, the PJQM2 lengths are in general larger (*right panel*).

The percentage of observations covered in each method, computed based on the concept of TEC, presented in Table 3. The corresponding TECs are also presented in the last column. Here we see that PJQM1 has the closest TEC to the nominal level, as it also supported by its wider intervals. On the other hand, LQMM also gives similar TEC but with shorter lengths. In this case LQMM overcomes PJQM1. PJQM2 gives the lowest TEC and hence only 3.87% of the new observations are covered by this method (LQMM can cover 6.29%, but higher TEC). Since the TECs are different between PJQM2 and the other two methods, we therefore cannot directly compare their IRI lengths, although PJQM2 lengths are larger than LQMMs.

This finding is slightly different with the observation in Fig. 5, where LQMM lengths tend to be larger. This might be because for Fig. 3, we simulated the datasets from normal distribution with normal subject-specific variances; the same assumption that has been applied in LQMM. Another example of IRI lengths comparison from non-Gaussian data can be consulted in Figure S3 and Table S2 in the *Supplementary Document*. We also show how the IRI lengths can affect the coverage in Figure S4, for all scenarios implemented in the simulation study. Despite the closer coverages to the true nominal level, LQMM still have some drawbacks; the LQMM method does not always achieve convergence, mostly because of a non-positive definite variance-covariance matrix of the random effects (see Figure S6 in the *Supplementary Document.*).

We have also evaluated the performance of the PJQM2 method by varying the shape and scale parameters a and b in the data generation, which is related to the subject-specific variances. The results are presented in Table 4. The number of subjects N, the number of repeated

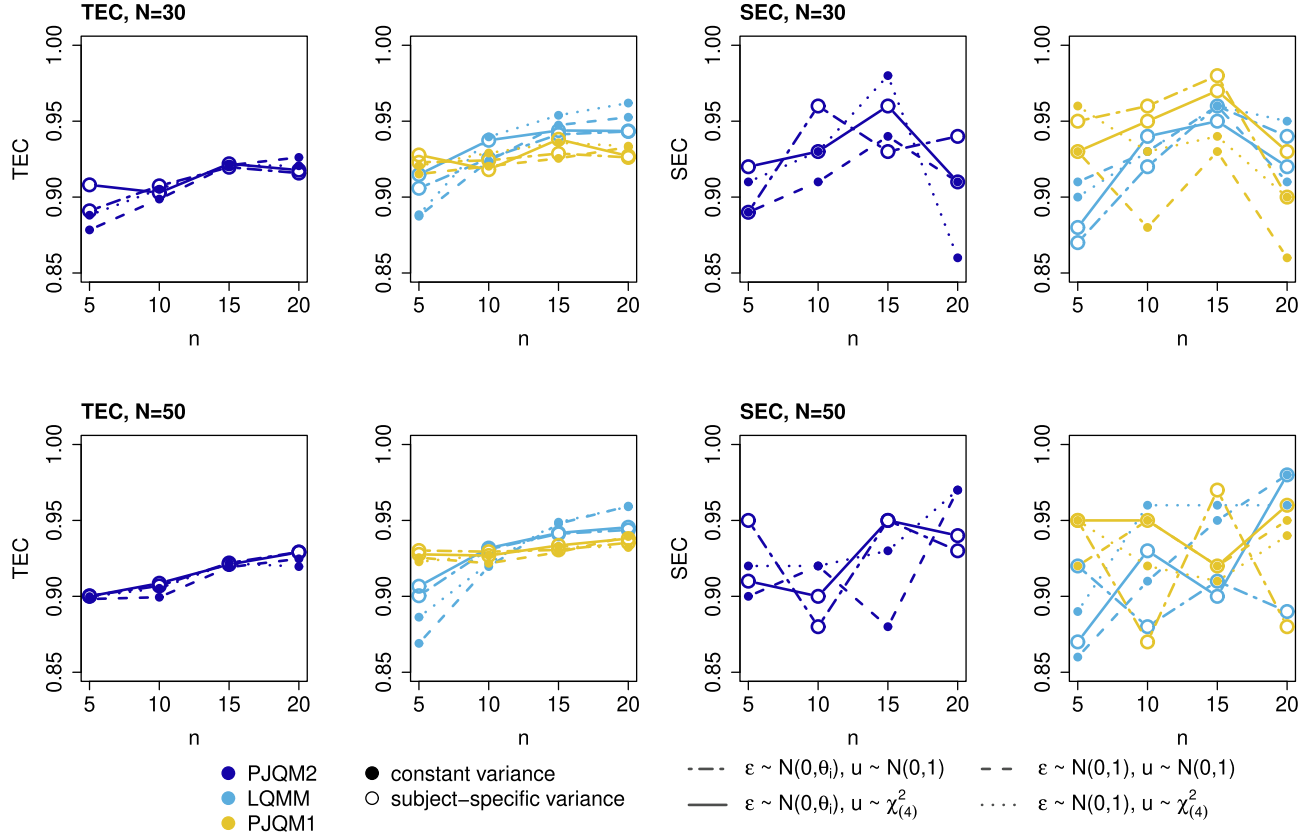

Fig. 2. Method performance of IRI computed using LQMM, PJQM2, and PJQM1 at the 95% nominal coverage. Data were generated from different error term and random effect distributions with either constant or subject-specific variance. The TEC and SEC were calculated for all scenarios of four different number of repeated measurements n in 30 and 50 subjects.

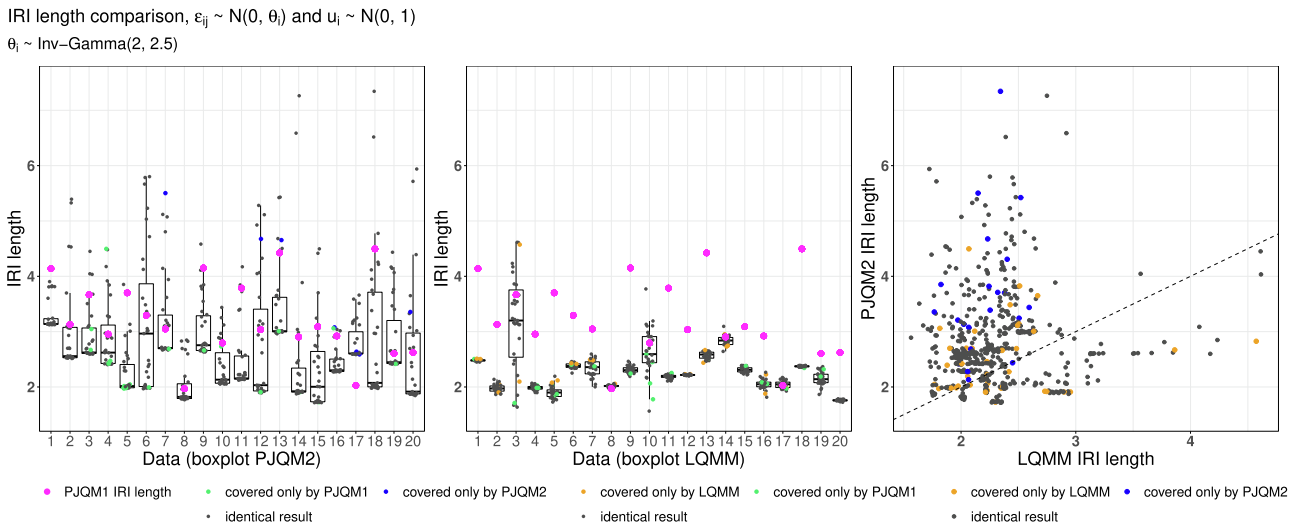

Fig. 3. The IRI length comparison in 20 simulated datasets. The IRI lengths of PJQM1 were overall larger than the lengths of PJQM2. The graph also gives information about when the IRI of PJQM2 covers the new observations (similar to the concept of TEC) but PJQM1/LQMM does not, and vice versa. All data were generated in the scenario of $N = 30$ and $n = 5$ with subject-specific variances ($\varepsilon_{ij} \sim \mathrm{N}(0, \theta_i)$ with θ_i ~Inverse-Gamma(2, 2.5); $u_i \sim \mathrm{N}(0, 1)$), at 95% nominal coverage.

Table 3
Confusion matrix of the percentage of observations covered by IRIs of one method, but not by the other. The IRIs were estimated in 20 simulated data generated in the scenario of $N = 30$ and $n = 5$ with subject-specific variances ($\varepsilon_{ij} \sim \mathrm{N}(0, \theta_i)$ with θ_i ~Inverse-Gamma(2, 2.5); $u_i \sim \mathrm{N}(0, 1)$). Example: 3.87% of the new observations in 20 simulated data are covered by PJQM2, but not by LQMM. The last column indicates the average of TEC by each method, for which it was also used to compute the percentage of observations coverage.

		Not covered by			TEC
		LQMM	PJQM1	PJQM2	
Covered by	LQMM		4.19%	6.29%	0.9145
	PJQM1	4.35%		3.71%	0.9161
	PJQM2	3.87%	1.13%		0.8903

Table 4
Performance of the PJQM2 method for different values of the parameters a and b of the inverse gamma distribution. The variation of the within-subject measurements as well as the IRI's lengths increase as the b parameter increases, while keeping the a parameter constant. All scenarios were evaluated at the 95% nominal coverage level in 30 subjects with 10 repeated measurements.

a	b	TEC	SEC
2	0.62	0.91	0.88
2	1.25	0.91	0.93
2	2.50	0.91	0.96
2.4	0.62	0.89	0.96
2.4	1.25	0.91	0.96
2.4	2.50	0.89	0.86
2.8	0.62	0.89	0.91
2.8	1.25	0.90	0.91
2.8	2.50	0.91	0.97

measurements within subjects n, and the nominal levels were fixed to 30, 10, and 95%, respectively. We observe that there is no strong impact of different a and b on the method's performance, except a slight increase of SEC in larger b values.

3.2. VITO IAM frontier data

Since all data comes from healthy subjects, the data can be used to investigate the performance of the LQMM, PJQM2, and PJQM1 methods. Results for the parametric method of Coşkun et al. [10] can be found in the *Supplementary Document*. As before, we will compute the empirical coverages (TEC and SEC) for each clinical test and metabolomics measurement, using the same procedure as described in Section 2.4.4. In this way, the results presented in this section can be considered as a realistic, empirical assessment of the coverages. At the same time we use the data to illustrate the use of the IRIs. We have calculated the 95% IRI using the LQMM, PJQM2 and PJQM1 methods for all eleven clinical tests and metabolomics measurements, measured in five time-points from bimonthly visits. In Fig. 4, we show those IRIs of creatinine in the metabolomics dataset. The graph shows that the IRIs vary between subjects and that the lengths of the intervals are generally smaller for the IRIs than for the published PRI, especially for PJQM1 and PJQM2, suggesting that IRIs give more precise information for individuals than the PRI. The LQMM, on the other hand, gives larger IRIs where the lengths are similar to the published PRI. This large intervals are not favourable, as they are less precise. The LQMM method estimation assumes a normal subject-specific effect, which may not be the case for this creatinine data, and hence overestimating the IRIs.

These IRIs from all the three methods were estimated based on five historical data of subjects, and should be used for interpreting the future measurement(s) i.e. the sixth measurement, as indicated in the figure. All of the future measurements are inside the PRI, but when we examine subject 25, the new measurement is now outside the IRI but it is still within the published PRI. A similar case occurs for subject 4, when the new measurement is at the IRI's borderline (from PJQM2) of the lower bound. In this case, the IRI can provide an early signal of potential abnormalities on creatinine level for this particular subject. Other examples of IRI implementation in the IAM Frontier metabolomics datasets can be consulted in Figure S5 and Figure S6 in the *Supplementary Document*.

Fig. 4 also illustrates that the lengths of the PJQM2 IRIs of some subjects are mostly smaller than for LQMM, while they are constant for PJQM1. The distributions of these lengths are shown as boxplots in Fig. 5. These boxplots illustrate three situations. For albumin, creatinine, HDL cholesterol and triglycerides, the lengths of the PJQM2 intervals are generally smaller than for LQMM. These wider lengths of LQMM result in a very high coverage, as reported in Table 5. Moreover, for creatinine and triglycerides, the lengths of the LQMM intervals do not have enough variation; this was caused by the convergence problem in the parameter estimation. However, although the lengths of the PJQM2 intervals are smaller, the method still can give coverages that are close to the nominal level. This supports the underlying idea of an IRI: it should

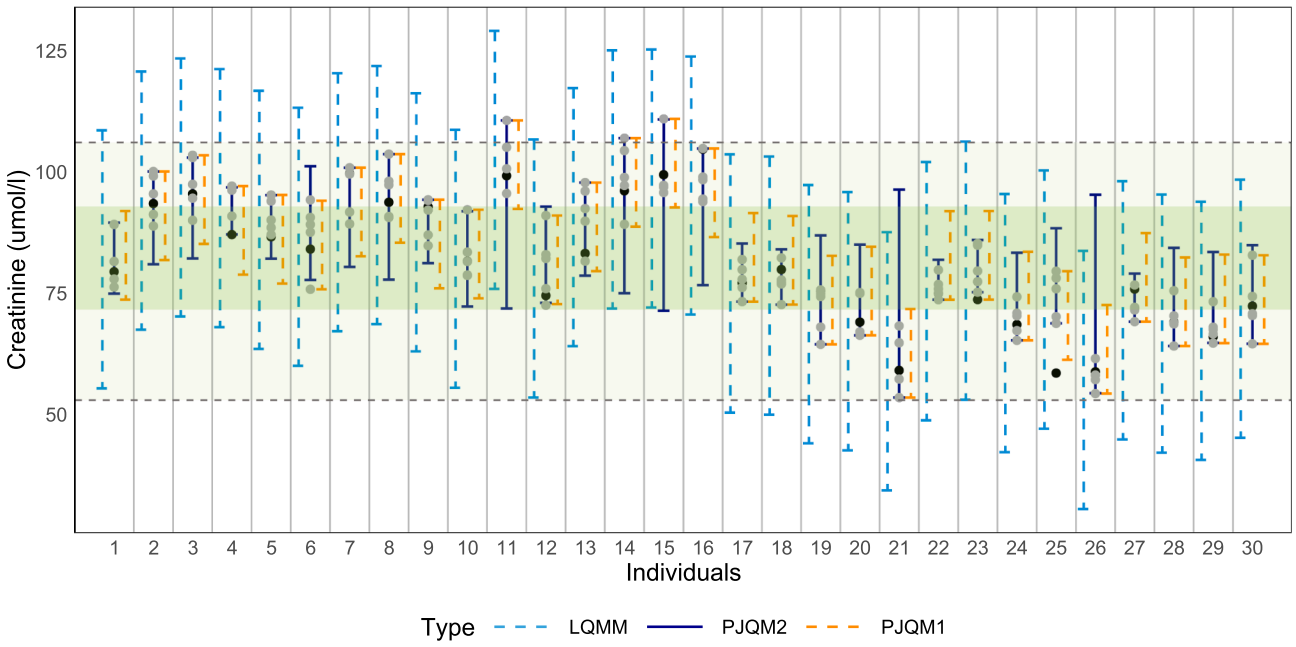

Fig. 4. The 95% IRIs computed using LQMM, PJQM2, and PJQM1 for all 30 subjects for creatinine in the VITO IAM Frontier metabolomics dataset. The gray dotted points indicate the historical data used for the IRI estimation and the black dotted points are the 6th (future) measurement. The light green transparent area indicates the published PRI [13]. The darker green transparent area is the 'average' PJQM2 i.e. when $u_i = 0$ and $z_i = 1$ in PJQM2. The order of individuals is randomised to ensure confidentiality.

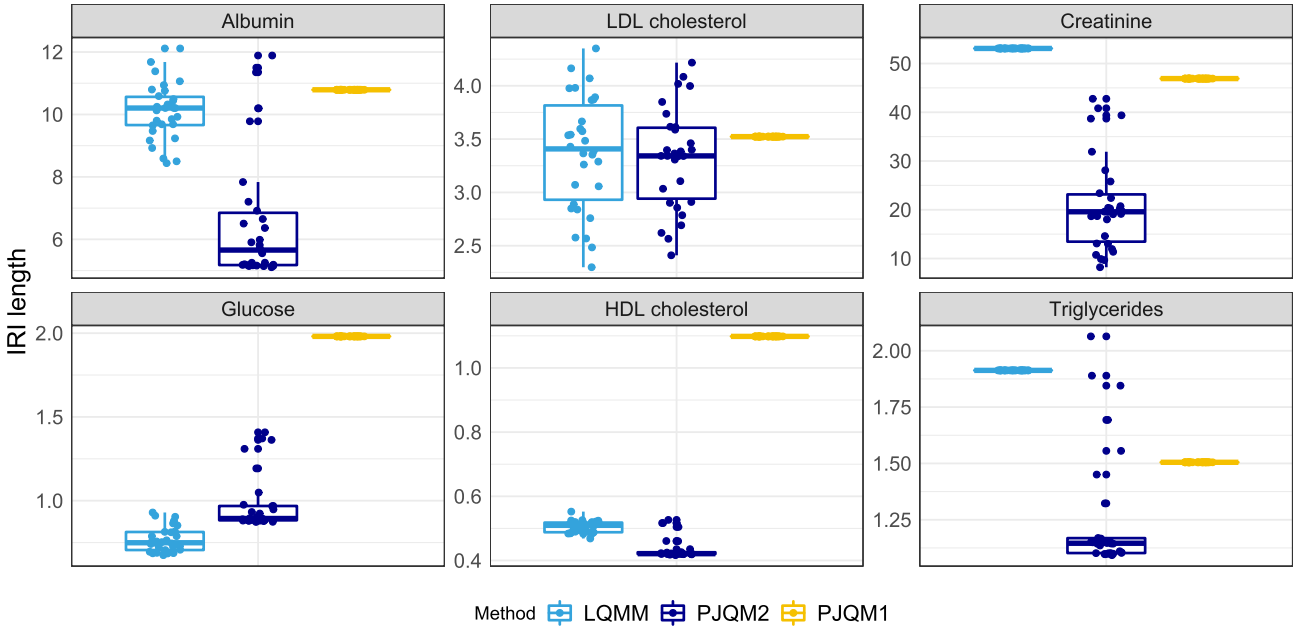

Fig. 5. The IRI length distributions of 6 metabolomics measurements measured from the VITO IAM Frontier. The PJQM2 method generally gives lower IRI length with smaller variability.

provide an individual-specific interval with correct coverages while keeping the length as small as possible. The IRI can thus be more personalised, while still having the correct protection against false positive results. LDL cholesterol is an example for which the length distributions are about the same, but only the empirical coverages of the PJQM2 intervals are close to the nominal level. Finally, for glucose, the lengths of PJQM2 are larger than those of LQMM, but PJQM2 still gives a greater TEC.

The TEC and SEC for other metabolomics measurements and clinical tests are presented in Table 5. A TEC and SEC as small as 50% and 57% are observed for LQMM, whereas they are rather consistent at 83–97% for PJQM2. Unsurprisingly, PJQM1 gives high TECs. This can be explained by the larger IRI lengths produced by PJQM1, as shown in Fig. 5, and therefore the new measurements are more likely fall within the IRIs and hence higher coverages.

We consider IRIs estimated using the PJQM2 method of creatinine and albumin computed from five historical data from the clinical and metabolomics datasets to discuss some of the features of the IRIs (Fig. 6).

Table 5
Benchmarking. SEC and TEC for the VITO IAM Frontier data for the 95% IRI. The values between 95% ± 2.5% are printed in bold face.

	LQMM		PJQM2		PJQM1	
	SEC	TEC	SEC	TEC	SEC	TEC
Albumin^{C}	1.000	1.000	0.867	0.867	0.600	1.000
Albumin^{M}	1.000	1.000	**0.967**	0.867	0.567	**0.933**
Apolipoprotein A1^{C}	0.567	0.500	0.833	0.700	0.867	**0.933**
Apolipoprotein A1^{M}	0.633	0.767	0.833	0.833	0.833	**0.933**
Apolipoprotein B^{C}	0.900	0.567	**0.933**	0.900	**0.967**	**0.933**
Apolipoprotein B^{M}	0.900	**0.933**	**0.933**	**0.967**	0.900	1.000
Ratio Apolipoprotein B/A1^{C}	0.700	1.000	**0.967**	0.900	0.900	**0.933**
Ratio Apolipoprotein B/A1^{M}	0.900	**0.967**	**0.967**	0.900	0.900	**0.967**
LDL cholesterol^{C}	0.900	**0.967**	**0.967**	0.900	**0.967**	**0.967**
LDL cholesterol^{M}	**0.933**	**0.967**	**0.967**	1.000	1.000	**0.967**
Creatinine^{C}	1.000	1.000	0.867	0.800	0.867	**0.933**
Creatinine^{M}	1.000	1.000	**0.933**	0.900	0.800	1.000
Glucose^{C}	0.900	1.000	**0.967**	**0.967**	0.700	1.000
Glucose^{M}	**0.967**	0.867	0.867	0.900	0.700	1.000
HDL cholesterol^{C}	0.900	**0.933**	0.833	0.833	0.900	0.900
HDL cholesterol^{M}	0.733	**0.967**	0.900	0.900	0.800	**0.967**
non-HDL cholesterol^{C}	0.767	0.733	**0.933**	**0.933**	0.900	**0.933**
non-HDL cholesterol^{M}	0.767	**0.967**	**0.967**	**0.967**	**0.967**	1.000
Total cholesterol^{C}	0.867	**0.967**	0.867	0.867	0.900	**0.933**
Total cholesterol^{M}	0.867	0.900	**0.967**	0.900	**0.967**	**0.933**
Triglyceride^{C}	0.833	**0.967**	**0.967**	**0.967**	**0.933**	**0.967**
Triglyceride^{M}	**0.933**	**0.967**	**0.967**	**0.967**	1.000	1.000

^{C}Clinical test; ^{M} Metabolomics measurement.

As before, the IRIs show between subject variability. While for creatinine most of the IRIs fall completely within the PRI, some subjects have IRIs that have one or two bounds outside of the PRI (e.g. subject 14 and 15). If it is indeed correct that all subjects in the IAM Frontier dataset are healthy, this is an illustration of a subject for whom large creatinine levels are normal. On the other hand, the new clinical test/measurement at the sixth time point is still within both the PRI and its IRI, hence the same conclusion. Had the new clinical test/measurement fallen outside the IRI but still within the PRI, e.g. in subject 13 and 25, this may perhaps be interpreted by the GP as an indication that this subject is at risk of impaired kidney function or kidney disease.

Still in Fig. 6 left panel, here we can also clearly see the effect of information-sharing, which is a desirable consequence of the estimation method and which is needed to allow for the IRI calculation based on only a few observations per subject. In particular, several subjects have observations that show rather small variability (e.g. subjects 21 and 26). However, their IRIs are much wider than suggested by their data. Note that their observations are not evenly distributed within their IRIs. This is a consequence of the information-sharing: some other subjects have much larger creatinine concentrations (e.g. subjects 11 and 15) with larger variability, and this also affects the IRIs of the other subjects. This not a problematic feature, because over all subjects the coverage is controlled. Moreover, the IRIs can also be seen as a compromise between the PRI (population interpretation) and the genuine individual RI that would be estimated from only the subject's data.

The IRIs for albumin show again between-subject variability. Here we see that the albumin concentrations are in general higher in clinical data than in metabolomics, hence the IRIs are also vary between them. The effect of information sharing is also present, such as in subject 26 and subject 30.

4. Discussion and conclusion

We have proposed three individual reference intervals estimation methods that are applicable for the personalised interpretation of clinical tests and metabolomics measurements. Linear Quantile Mixed Models (LQMM) serve as a basis for the first method for constructing the IRI. The variability between the IRI motivates the need for subject-specific interpretation of reference intervals [2]. The Penalised Joint Quantile Model with two subject-specific effects (PJQM2) and its penalised estimation procedure do not require strong distributional assumptions, in contrast to the LQMM method. Due to the joint modelling of the lower and upper quantiles, and the introduction of penalty parameters, the subject-specific parameters (that specify the position and the length of the IRI) are estimated using all available data. This can also be interpreted as an example of information sharing: even though each subject may only contribute a small number of measurements that is insufficient for the nonparametric estimation of a reference interval, information is shared among subjects, allowing for the estimation of the intervals. We have also described the Penalised Joint Quantile Model with one subject-specific effect (PJQM1), which also incorporates a penalty parameter and the same selection procedure as in the PJQM2 method is applied. There is only one subject-specific parameter in the

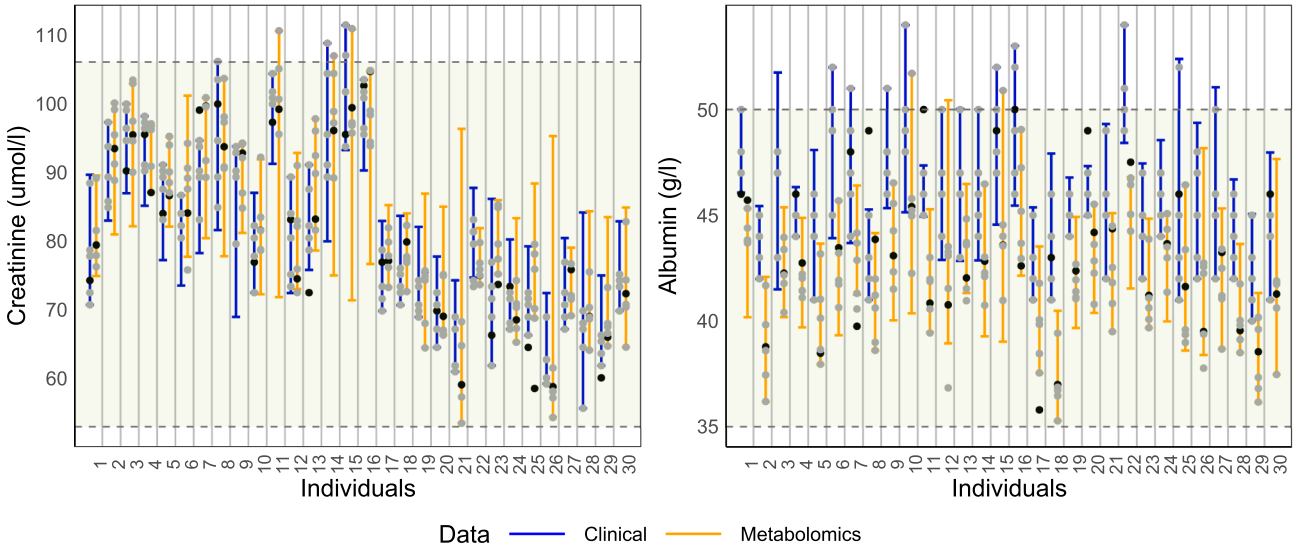

Fig. 6. The IRIs for all subjects in the VITO IAM Frontier dataset, estimated using the PJQM2 method. The gray dotted points indicate the historical data used for the IRI estimation and the black dotted points are the 6th (future) measurement. The green transparent area indicates the published PRI [13]. The order of individuals is randomised and different between creatinine and albumin.

model and it is shared between the upper and the lower quantiles, making the length of the IRIs constant over all subjects.

Our simulation study demonstrates that for various sample sizes and numbers of repeated measurements, the IRI estimated by PJQM1 always give time empirical coverage (TEC) values close to the true nominal coverage level, often outperforming LQMM and PJQM2. However, the high coverage values of the PJQM1 intervals are related to greater lengths of its IRIs. It also has a feature of providing a constant length for all subject, leading into a less personalised interpretation. On the other hand, LQMM gives lower TEC and SEC values than PJQM1, and it does not perform well in a dataset with rather small numbers of repeated measurements. Therefore, the robustness of our new PJQM2 method in estimating the IRI is suggested: it gives consistent empirical coverages that are not strongly affected by the sample size, the number of repeated measurements, and the shape of the data distributions.

In current practice, clinical decisions are usually binary: if the observation is within the PRI, it is considered normal. If the observation is outside the PRI, an action is usually taken. The IRI range extends this interpretation by providing a personal context to the PRI. The IRI also offers additional information to the PRI for the preventive purpose of one's health. For example, if an individual's current clinical test of a particular marker is outside the IRI but still within the PRI, the individual is likely in an early phase of developing a disease, w.r.t. that clinical marker. The health condition of that individual might still be favorable at the time when the test was taken, but the IRI suggests a potential deviation from its own healthy range. Hence, a preventive action should be taken.

The proposed methods are based on quantile models and therefore inherit the flexibility of a statistical modelling framework. An extension of the models is also being studied for incorporating baseline characteristics of the subjects, such as age and sex. This is important as individual's characteristics are generally different between males and females, and it is known that clinical tests may evolve over time (e.g. age), even if the individual is in a healthy state. Combining this information would result in a more precise IRI as well as increasing its performance, both in detecting the abnormalities in future measurements and in a new subject. The sex-specific effect can actually be seen in Fig. 6, where the IRIs of females (subject 16–20) are affected by the high creatinine levels in male subjects (subject 1–15). When the model is extended by correcting the sex effect, we will see a shorter IRIs in both males and females. This extension would be a topic for further research.

Future studies with longitudinal data including both healthy and diseased populations could give valuable insights into the clinical utility of IRIs. Furthermore, currently it is a challenge to collect multiple data points from a single individual (which is necessary for calculating IRIs) where his/her data is scattered across various hospitals, institutions and e-health records due to technical difficulties and privacy concerns [18]. Cloud-based Linked Data solutions are being developed at the European level to address these challenges [19], hence enabling data coverage at the individual level.

We believe that when the IRI is widely used in daily clinical practice for interpretation of the results complementary to PRIs, it will be immensely beneficial to extend the use of clinical, metabolomics, and proteomics markers to precision (and preventive) health.

Data Availability

The data that supports the simulation findings are reproducible. The codes to generate the data and produce the results are available online at https://github.com/murihpusparum/PenalizedJQM and in the Supplementary Document. The VITO IAM Frontier data is subject to data protection and privacy of data subjects. It can be made available upon request.

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.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jbi.2022.104111.

References

[1] R. Moore, An historical perspective on the clinical diagnostic laboratory, in: J. Fagerberg, D.C. Mowery, R.R. Nelson, W.B. Coleman, G.J. Tsongalis (Eds.), Molecular Diagnostics, Humana Press, New Jersey, 2006, pp. 3–10. https://doi.org/10.1385/1-59259-928-1:003.

[2] M. Pusparum, G. Ertaylan, O. Thas, From population to subject-specific reference intervals, Comput. Sci. ICCS 2020 (12140) (2020) 468–482, https://doi.org/10.1007/978-3-030-50423-6_35.

[3] Y. Ozarda, K. Sikaris, T. Streichert, J. Macri, on behalf of IFCC Committee on Reference intervals, (C-RIDL), D.L. Distinguishing reference intervals and clinical decision limits – a review by the ifcc committee on reference intervals and decision limits, Crit. Rev. Clin. Lab. Sci. 55 (6) (2018) 420–431, https://doi.org/10.1080/10408363.2018.1482256.

[4] H. Solberg, Approved recommendation (1987) on the theory of reference values. Part 5: Statistical treatment of collected reference value. determination of reference limit, Clinica Chim. Acta 170 (1987) S13–S32. https://doi.org/10.1016/0009-8981(87)90151-3.

[5] E.M. Wright, A comparison of statistical methods for age-related reference intervals, J. R. Statist. Soc. Ser. A (Statist. Soc.) 160 (1) (1997) 47–69, https://doi.org/10.1111/1467-985X.00045.

[6] S. Wellek, K. Lackner, C. Jennen-Steinmetz, et al., Determination of reference limits: statistical concepts and tools for sample size calculation, Clin. Chem. Lab. Med. 52(12) (2014) 1685–1694. https://doi.org/10.1515/cclm-2014-0226.

[7] S. Poole, L.F. Schroeder, N. Shah, An unsupervised learning method to identify reference intervals from a clinical database, J. Biomed. Informat. 59 (2016) 276–284, https://doi.org/10.1016/j.jbi.2015.12.010.

[8] H. Solberg, Refval: a program implementing the recommendations of the international federation of clinical chemistry on the statistical treatment of reference values, Comput. Methods Programs Biomed. 48 (1995) 247–256. https://doi.org/10.1016/0169-2607(95)01697-x.

[9] H. Solberg, R. Grasbeck, Reference values. Advances, Clin. Chem. 27 (1989) 1–79, https://doi.org/10.1016/s0065-2423(08)60181-x.

[10] A. Coskun, S. Sandberg, I. Unsal, C. Cavusoglu, M. Serteser, M. Kilercik, et al., Personalized reference intervals in laboratory medicine: A new model based on within-subject biological variation, Clin. Chem. 67 (2) (2021) 374–384, https://doi.org/10.1093/clinchem/hvaa233.

[11] R. Koenker, Quantile regression for longitudinal data, J. Multivariate Anal. 91 (1) (2004) 74–89, https://doi.org/10.1016/j.jmva.2004.05.006.

[12] I AM Frontier study - VITO, Belgium. https://iammyhealth.eu/en/i-am-frontier (accessed: 03 December 2021).

[13] M. Laposata, Laboratory medicine: The diagnosis of disease in the clinical laboratory, second ed., McGraw-Hill Education, New York, 2014.

[14] M. Geraci, M. Bottai, Quantile regression for longitudinal data using the asymmetric laplace distribution, Biostatistics 8 (1) (2007) 140–154, https://doi.org/10.1093/biostatistics/kxj039.

[15] M. Geraci, M. Bottai, Linear quantile mixed models, Statist. Comput. 24 (2014) 461–479, https://doi.org/10.1007/s11222-013-9381-9.

[16] M. Geraci, Linear quantile mixed models: The lqmm package for laplace quantile regression, J. Statist. Softw. 57 (13) (2014) 1–29, https://doi.org/10.18637/jss.v057.i13.

[17] rqpd: Regression quantiles for panel data. R package version 0.6/r10. https://R-Forge.R-project.org/projects/rqpd/ (accessed: 03 December 2021).

[18] K. Boonen, K. Hens, G. Menschaert, et al., Beyond genes: Re-identifiability of proteomic data and its implications for personalized medicine, Genes 10(9) (2019) 682. https://doi.org/10.3390/genes10090682.

[19] A. Zaveri, G. Ertaylan, Linked data for life sciences, Algorithms 10 (4) (2017) 126, https://doi.org/10.3390/a10040126.