







# RIbench: A Proposed Benchmark for the Standardized Evaluation of Indirect Methods for Reference Interval Estimation

Tatjana Ammer <sup>a,b,†</sup> André Schützenmeister <sup>b</sup> Hans-Ulrich Prokosch <sup>a</sup> Jakob Zierk <sup>c,d</sup>  
Christopher M. Rank <sup>b</sup> and Manfred Rauh <sup>c,\*</sup>

**BACKGROUND:** Indirect methods leverage real-world data for the estimation of reference intervals. These constitute an active field of research, and several methods have been developed recently. So far, no standardized tool for evaluation and comparison of indirect methods exists.

**METHODS:** We provide RIbench, a benchmarking suite for quantitative evaluation of any existing or novel indirect method. The benchmark contains simulated test sets for 10 biomarkers mimicking routine measurements of a mixed distribution of non-pathological (reference) values and pathological values. The non-pathological distributions represent 4 common distribution types: normal, skewed, heavily skewed, and skewed-and-shifted. To identify strengths and weaknesses of indirect methods, test sets have varying sample sizes and pathological distributions differ in location, extent of overlap, and fraction. For performance evaluation, we use an overall benchmark score and sub-scores derived from absolute z-score deviations between estimated and true reference limits. We illustrate the application of RIbench by evaluating and comparing the Hoffmann method and 4 modern indirect methods – TML (Truncated-Maximum-Likelihood), kosmic, TMC (Truncated-Minimum-Chi-Square), and refineR – against one another and against a nonparametric direct method ( $n = 120$ ).

**RESULTS:** For the modern indirect methods, pathological fraction and sample size had a strong influence on the results: With a pathological fraction up to 20% and a minimum sample size of 5000, most methods achieved results comparable or superior to the direct method.

**CONCLUSIONS:** We present RIbench, an open-source R-package, for the systematic evaluation of existing and

novel indirect methods. RIbench can serve as a tool for enhancement of indirect methods, improving the estimation of reference intervals.

## Introduction

Big data has become an integral part of medicine, including basic and clinical research to improve clinical decision-making. Laboratory test results performed during routine patient care, the so-called real-world data (RWD), are one important example of big data in healthcare (1, 2). Establishing reference intervals (RI) is a prominent application where RWD has the potential to increase the precision of reference limits, leading to improved interpretation of test results (especially in populations where conventional methods are limited, e.g., children or the elderly) (2–4).

To establish RIs, CLSI currently recommends conducting a direct RI study using samples collected from at least 120 subjects from a predefined reference population (5). Indirect methods use data-mining procedures and statistical algorithms in combination with RWD, and estimate RIs by identifying the non-pathological distribution (4, 6–11).

Several indirect methods have been developed and are already widely implemented in different settings (12–16). Although their performance has been analyzed individually (17–19), these studies are not directly comparable, as the algorithm performance depends on characteristics of the input dataset. To overcome this limitation, and in accordance with the ongoing activities for global standardization and harmonization in laboratory medicine, a “benchmark” (20) for the standardized evaluation of indirect methods is required.

<sup>a</sup>Friedrich-Alexander-Universität Erlangen-Nürnberg, Chair of Medical Informatics, Erlangen, Germany; <sup>b</sup>Roche Diagnostics GmbH, Biostatistics & Data Science, Penzberg, Germany; <sup>c</sup>Universitätsklinikum Erlangen, Department of Pediatrics and Adolescent Medicine, Erlangen, Germany; <sup>d</sup>Universitätsklinikum Erlangen, Center of Medical Information and Communication Technology, Erlangen, Germany.

\*Address correspondence to this author at: Loschgestr. 15, 91054 Erlangen, Germany. Fax +49-9131-85-33714; e-mail manfred.rauh@uk-erlangen.de.

<sup>†</sup>The present work was performed in partial fulfillment of the requirements for obtaining the degree “Dr. rer. biol. hum.”  
Received April 14, 2022; accepted July 12, 2022.  
<https://doi.org/10.1093/clinchem/hvac142>

The concept of benchmarking is common in computer science fields, e.g., for the comparative evaluation of computer systems, compilers, or databases (20). For quality assurance in laboratory medicine, for instance, ring trials are used to evaluate, ensure, and compare the analytical performance of different assays. Likewise, benchmarking has the goal to evaluate and compare different methods with respect to specified performance measures based on a pre-defined set of tests. Thus, it enables a systematic, reproducible, transparent, and objective evaluation. Moreover, it serves as a tool to identify the best method for a specific use case, as well as to enhance existing methods (20).

We apply this concept to indirect methods and provide RIbench, an R-based open-source benchmark. This benchmark enables performance evaluation of any existing or novel indirect method in estimating RIs from datasets consisting of a mixture of non-pathological (reference) and pathological samples.

Furthermore, with RIbench we aim to answer questions that frequently arise when discussing the use of indirect methods in laboratory practice (6, 7, 21): How is the performance influenced by the distribution type of the non-pathological distribution? What is the impact of the sample size on the performance? Up to what fraction of pathological samples or up to what amount of overlap between non-pathological and pathological distributions do the algorithms report reliable results?

As a proof of principle, we evaluate 5 different indirect methods: the Hoffmann method (22), and 4 modern methods: TML (Truncated Maximum Likelihood) (8, 23–26), TMC (Truncated Minimum Chi-square) (9), kosmic (10), and refineR (11).

## Materials and Methods

Our benchmarking suite, RIbench (version 1.0), comprises a large number of simulated biomarker test sets, which enables quantitative performance evaluation of indirect methods, as the true reference limits are known. These correspond to Box–Cox-transformed and possibly shifted sample quantiles of normal distributions.

### DEFINITION OF BENCHMARKING SUITE TEST SETS

The benchmark test sets are designed to mimic RWD of well-known biomarkers. They are derived from analysis of data sets with several thousand real routine measurements and are simulated in accordance with RIs appropriate for the respective biomarker. To get a representative set of the different non-pathological distribution types occurring in laboratory practice, we simulated non-pathological data for 10 different biomarkers (Fig. 1, Table 1): (approximately) normal distributions for hemoglobin (Hb), calcium (Ca), and free thyroxine (FT4), skewed

distributions for aspartate transaminase (AST), lactate (LACT), and gamma-glutamyltransferase (GGT), heavily skewed distributions for thyroid-stimulating hormone (TSH), immunoglobulin E (IgE), C-reactive protein (CRP), and a skewed-and-shifted distribution for lactate dehydrogenase (LDH). The distributions representing these non-pathological samples are based on 1- or 2-parameter Box–Cox-transformed normal distributions (29) (see [Supplemental Methods](#) in the online Data Supplement). In a real-world scenario, these unimodal distributions would require appropriate physiological partitioning and would contain only one measurement per subject. The classification of the biomarkers to the different distribution types is based on the ratio of distances between the mode and the lower, and the upper reference limit (Eq. 1):

$$\text{Distance ratio} = \frac{|LRL - mode|}{|URL - mode|} \quad (1)$$

where *LRL* is the lower reference limit, *URL* is the upper reference limit, and *mode* is the mode of the underlying distribution. Biomarkers with a distance ratio  $\geq 0.75$  are considered normally distributed, while those with a ratio between 0.25 and 0.75 are considered skewed, and every analyte with a ratio  $< 0.25$  is assigned to the heavily skewed category.

Distributions representing pathological samples are simulated using normal distributions and are added to the left and the right side of the non-pathological distribution. To ensure a broad variety of biomarker test sets, the following parameters are varied (for a comprehensive overview of all parameters see [Table 1](#) and the accompanying R-package):

- (a) Sample size:  $n = 1000$ ;  $n = 5000$ ;  $n = 50\,000$ ;  $n = 500\,000$ .
- (b) Overall pathological fraction: 0%, 5%, 10%, 20%, 30%, 40%, 50%, 60%.
- (c) For each biomarker, 2 different scenarios are defined by varying the following parameters:
  - (i) Ratio between the fraction of “left” and “right” pathological distribution.
  - (ii) Location and width of pathological distributions.
  - (iii) Overlap between pathological and non-pathological distributions (3 levels of complexity, covering small, medium, and high overlap).

The overlap between the non-pathological and pathological distribution is defined as overlap on the concentration scale between the 2.5th and 97.5th percentiles of both distributions ([Supplemental Methods and Supplemental Eq. 1](#)). To simulate inconsistencies occurring in RWD, we added a background “noise” distribution (uniform distribution, fraction of 0.1%, [Table 1](#)). The algorithms to be tested do not receive

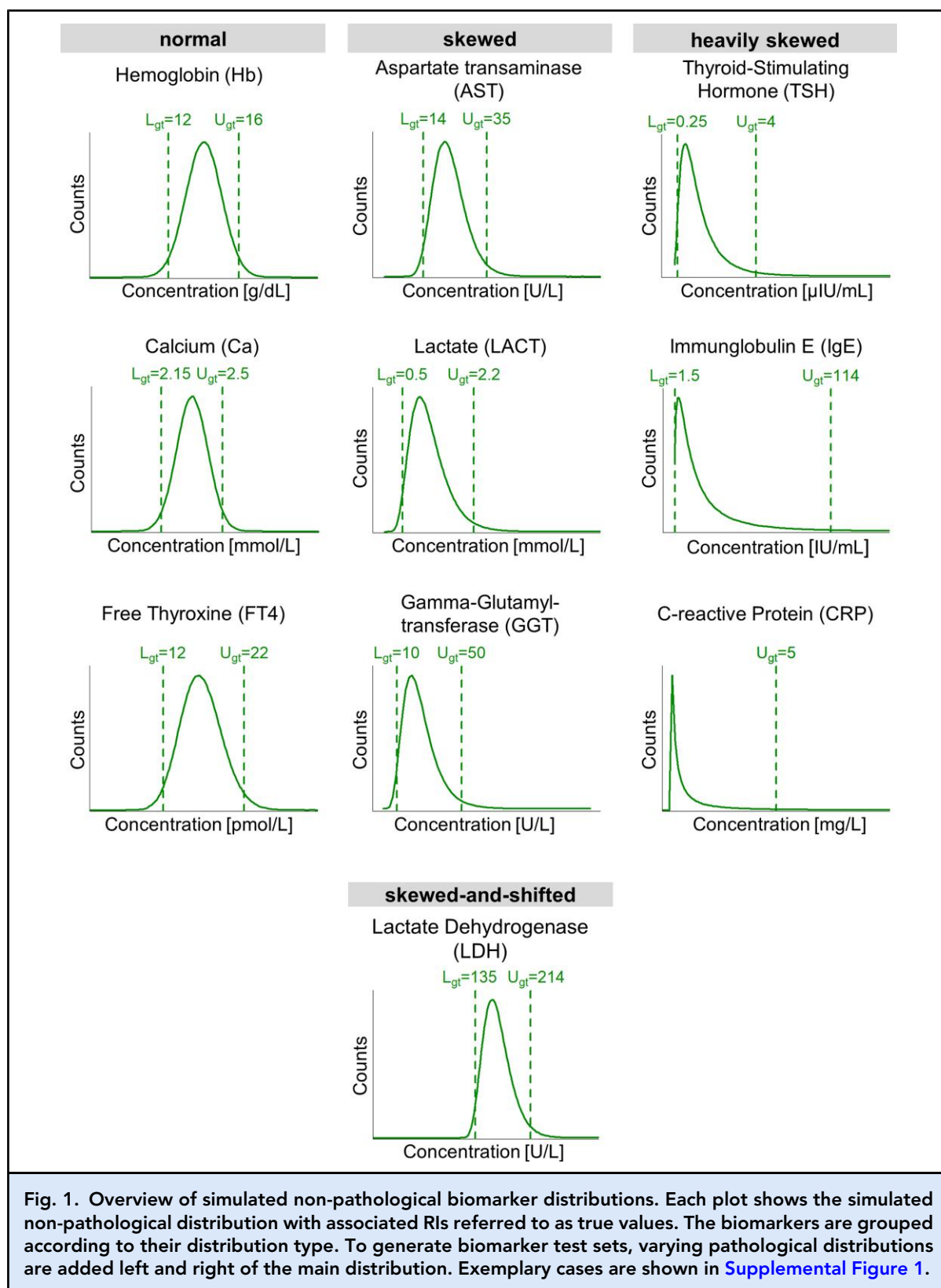


Table 1. Comprehensive overview of simulated biomarkers <sup>a</sup> .																		
Biomarker <sup>b</sup>	Non-pathological distribution				Pathological distribution								Background		Unit	Decimal points		
	Type (distance ratio)	RIs	μ	σ	λ	S	Ratio left: right		Left			Right						
							μ	σ	μ	σ	x-ov [%]	μ	σ	x-ov [%]				
Hb	Normal (1.00)	12.0-16.0 <sup>c</sup>	13	1.020427	1	0	1:1	10.18	1.03	5	17.82	1.03	5	0	160	g/dL	1	
								10.98	1.03	25	17.02	1.03	25					
								11.78	1.03	45	16.22	1.03	45					
								8.944	1.6	2	18.08	1.1	2					
								9.644	1.6	20	17.36	1.1	20					
							3:1	10.66	1.6	45	16.36	1.1	45					
								1.936	0.118	5	2.677	0.099	5					
								2.006	0.118	25	2.607	0.099	25					
								2.077	0.118	45	2.537	0.099	45					
								1.663	0.254	3	2.802	0.159	3					
Ca	Normal (0.94)	2.15-2.50 <sup>d</sup>	1.3254	0.0894	1	0	2:1	1.663	0.254	3	2.802	0.159	3	0	25	mmol/L	2	
								1.722	0.254	20	2.742	0.159	20					
								1.81	0.254	45	2.655	0.159	45					
								8.104	2.498	10	27.86	3.498	10					
								9.604	2.498	25	26.36	3.498	25					
FT4	Normal (0.82)	12.0-22.0 <sup>d</sup>	5.47089	0.5171	0.432	0	1:1	11.6	2.498	45	24.36	3.498	45	0	220	pmol/L	1	
								5.184	3.988	10	29.6	4.388	10					
								6.684	3.988	25	28.1	4.388	25					
								8.684	3.988	45	26.1	4.388	45					
								5.152	5.05	5	42.87	4.55	5					
AST	Skewed (0.50)	14-35 (27)	3.09721	0.233755	0	0	1:1	8.302	5.05	25	39.72	4.55	25	0	350	U/L	0	
								12.5	5.05	45	35.52	4.55	45					
								1.2	-2.128	8.55	3	47.99	6.95					3
								1.2	1.442	8.55	20	44.42	6.95					20
								1.2	5.643	8.55	40	38.12	6.95					50
Continued																		

Table 1. (continued)

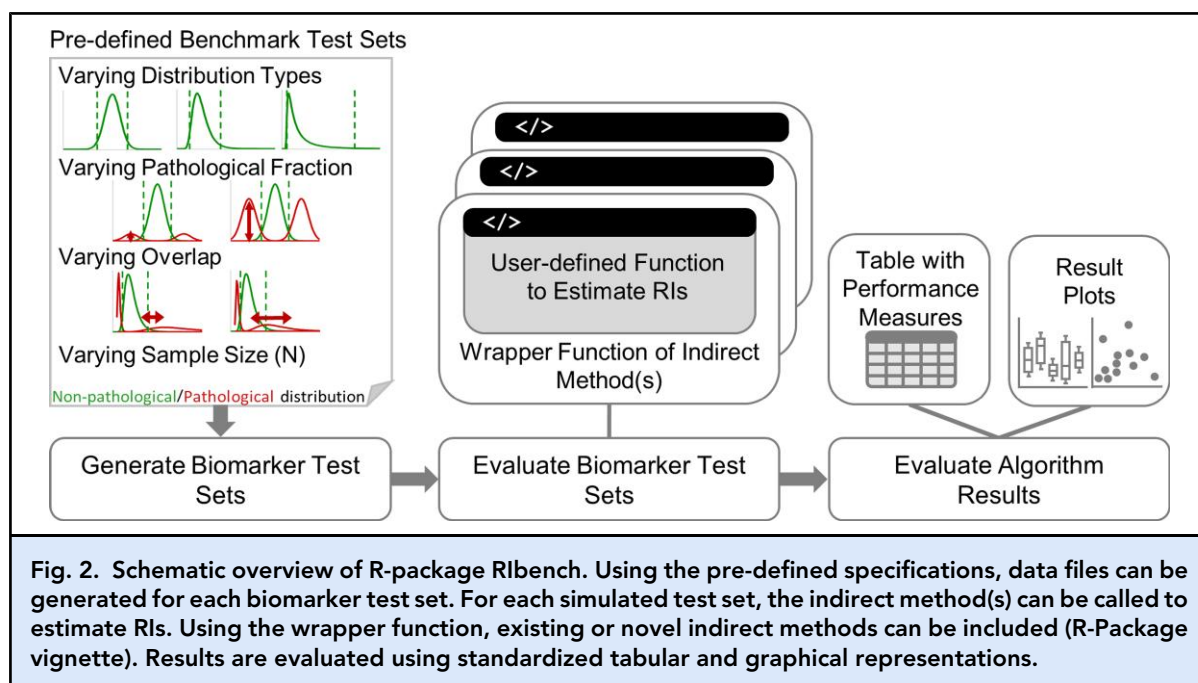
Table 1. (continued)																		
Biomarker <sup>b</sup>	Non-pathological distribution				Pathological distribution								Background		Unit	Decimal points		
	Type (distance ratio)	RIs	μ	σ	λ	S	Ratio left: right	Left				Right						
								μ	σ	x-ov [%]	μ	σ	x-ov [%]	Min			Max	
LACT	Skewed (0.31)	0.5-2.2 <sup>d</sup>	0.048	0.378	0	0	1:1	-0.1988	0.4	5	3.096	0.5	5	0	22	mmol/L	1	
							1:1	-0.1137	0.4	10	2.501	0.5	40					
							1:1	0.05627	0.4	20	2.160	0.5	60					
							1:3	-1.017	0.8	3	4.383	1.2	10					
							1:3	-0.9317	0.8	8	3.873	1.2	40					
							1:3	-0.8127	0.8	15	3.192	1.2	80					
							1:1	-1.72	7	5	65.64	9	5					
							1:1	0.2803	7	10	51.64	9	40					
GGT	Skewed (0.29)	10-50 <sup>c</sup>	3.107304	0.4105784	0	0	1:1	4.28	7	20	43.64	9	60	0	500	U/L	0	
							1:3	-5.64	9	5	71.52	12	5					
							1:3	-1.64	9	15	59.52	12	35					
							1:3	0.3603	9	20	45.52	12	70					
							1:1	-0.655	0.5	2	5.772	1	5					
							1:1	-0.355	0.5	10	4.46	1	40					
							1:1	0.01999	0.5	20	3.71	1	60					
							1:3	-1.635	1	2	7.732	2	5					
TSH	Heavily skewed (0.11)	0.25-4.00 <sup>c</sup>	-5.506435 × 10 <sup>-6</sup>	0.7073083	0	0	1:3	-1.335	1	10	6.795	2	30	0	40	μU/mL	2	
							1:3	-0.96	1	20	5.67	2	60					

Continued

Continued

Table 1. (continued)																	
Biomarker <sup>b</sup>	Type (distance ratio)	Non-pathological distribution					Pathological distribution						Background		Unit	Decimal points	
		RIs	$\mu$	$\sigma$	$\lambda$	S	Ratio left: right	Left			Right			Min			Max
								$\mu$	$\sigma$	x-ov [%]	$\mu$	$\sigma$	x-ov [%]				
IgE	Heavily skewed (0.02)	1.5-114 (28)	2.571	1.104715	0	0	1:1	-19.77	12	2	153.5	23	5	0	1140	IU/mL	0
								-16.39	12	5	102.8	23	40				
								-10.77	12	10	80.33	23	60				
								-12.47	10	5	182.9	38	5				
								-9.099	10	8	149.1	38	35				
CRP	Heavily skewed (0.01)	< 5.0 <sup>d</sup>	-1.1237	1.6617826	0	0	0:1	0	0	0	11.1	2	15	0	50	mg/L	1
								0	0	0	9.412	2	35				
								0	0	0	6.462	2	70				
								0	0	0	15.02	4	15				
								0	0	0	12.91	4	40				
LDH	Skewed + shifted (0.44)	135-214 <sup>d</sup>	4.73445	0.38824	0.0613	100	1:1	99.78	20	5	253.2	22	5	0	2250	U/L	0
								115.6	20	25	237.4	22	25				
								127.4	20	40	217.6	22	50				
								78.6	30	3	309.6	50	3				
								88.07	30	15	300.2	50	15				
						1:2	103.9	30	35	268.6	50	55					
<sup>a</sup> The table shows the biomarkers with distribution type, RIs, parameters used for the simulations, unit, and decimal points used for rounding the simulated data. For each simulated biomarker, two different scenarios for the pathological distributions are specified by mean ( $\mu$ ), standard deviation ( $\sigma$ ), fraction of left and right distributions. The mean directly influences the x-overlap (x-ov) between the non-pathological and pathological distribution. Thus, each line in the table defines a combination of the non-pathological distribution and the location of the pathological distributions. Additionally, the parameters used for simulating the uniform, background "noise" distribution are defined by specifying the minimum (Min) and maximum (Max). To create all test set combinations, sample size and pathological fractions are varied (see Materials and Methods).																	
<sup>b</sup> Biomarker abbreviations: Hemoglobin (Hb), Calcium (Ca), Free Thyroxine (FT4), Aspartate Transaminase (AST), Lactate (LACT), Gamma-Glutamyltransferase (GGT), Thyroid-Stimulating Hormone (TSH), Immunoglobulin E (IgE), C-reactive Protein (CRP), Lactate Dehydrogenase (LDH).																	
<sup>c</sup> Adapted from Zierk et al. (10).																	
<sup>d</sup> Extracted from method sheets provided by Roche Diagnostics GmbH, Mannheim.																	





any information about the parameters, or the pathological or non-pathological status of individual data points. One exception here is the variant of the Hoffmann method (with transformation), where the true Box–Cox power parameter is provided upfront.

Examples for a least, intermediate, and most challenging case for each biomarker are shown in online [Supplemental Fig. 1](#). The least complex case is defined as the one with the lowest pathological fraction and smallest overlap, the intermediate case as the one with a medium pathological fraction and a medium overlap, and the most challenging case is the one with the highest pathological fraction and largest overlap.

Overall, the benchmarking suite contains 5760 simulated test sets, 576 per biomarker. To analyze the computation time, we defined a fixed subset of test sets comprising 50 sets per biomarker.

To mimic RWD occurring in laboratory practice, we used a fixed number of decimal places, commonly used in laboratory practice ([Table 1](#)). Rlbench additionally provides the option to adapt the number of decimal places to evaluate the impact on performance when using input data with a different numeric precision ([7](#)).

#### IMPLEMENTATION — R PACKAGE RIBENCH

The presented benchmarking suite is implemented and provided as an open source R-package ([30](#)) on CRAN (<https://CRAN.R-project.org/package=Rlbench>) ([Fig. 2](#)). Rlbench enables integration and evaluation of any existing or novel indirect method (see example in R-package vignette). The simulated test sets are generated

using pre-specified initialization seeds for the random number generator to ensure reproducibility and objective comparison of results obtained by different users. Furthermore, the package provides a convenience function with options for analyzing a single (simulated) biomarker, a subset, e.g., for a specific distribution type, or all biomarkers. To evaluate the performance of an indirect method, different functions for tabular and graphical representation of the results are provided ([Fig. 2](#)).

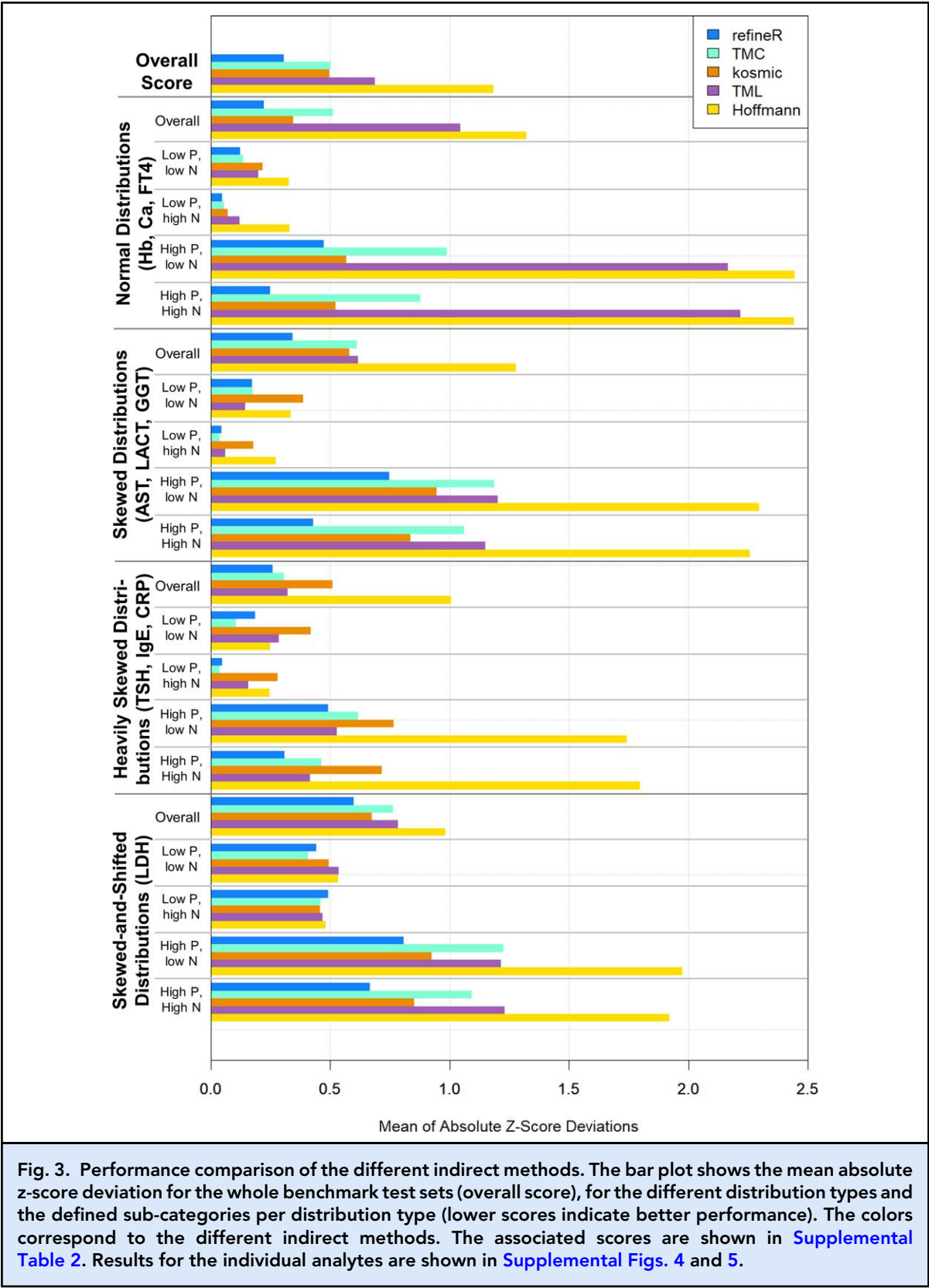
#### EVALUATION AND PERFORMANCE MEASURE

Because we use simulated test sets, estimates of indirect methods can be compared to the true values, enabling a quantitative performance evaluation. We use the absolute z-score deviation ([Eq. 2](#)) as a standardized performance measure. As this measure is independent of the skewness and concentration range of the non-pathological distribution, it enables the comparison and combination of results of different biomarkers:

$$\text{z-score deviation } zDevAbs = |z_{Est,i} - z_{T,i}| \quad (2)$$

$$\text{with } z_{Est,i} = \frac{\text{BoxCox}(RL_{Est,i} - S_T, \lambda_T) - \mu_T}{\sigma_T} \quad (3)$$

where  $\lambda_T$ ,  $S_T$ ,  $\mu_T$ ,  $\sigma_T$  are true power parameter, shift, mean, and standard deviation, respectively;  $RL_{Est,i}$  is the reference limit estimated by the indirect method with  $i$  specifying either the upper or lower limit;  $z_{Est,i}$  or  $z_{T,i}$  describes the computed z-score for the estimated and the true reference limits.





The z-score deviation (Eq. 2) represents the error on the standard deviation scale. Thus, the estimated RLs are first transformed into the standard deviation domain ( $z_{Est,i}$ ) (Eq. 3) and then the difference compared to the true z-scores ( $z_{T,i}$ ) is computed (i.e.,  $-1.96$  and  $1.96$  for the 2.5th and 97.5th percentiles) (Eq. 3). This absolute z-score deviation is calculated for both the lower and the upper reference limit, and the mean of these is computed as the collective deviation for each simulated test set.

To compare different methods on a global scale, one overall benchmark score per algorithm is calculated as the mean of all test set-specific deviations. To make the score robust against the influence of extremely large deviations, results with an absolute z-score deviation  $>5$  are considered as being implausible and excluded from the overall score. To have a fair comparison, the percentage of these implausible results is reported. Additionally, a failure rate is documented, capturing the amount of test sets for which a method fails to report a result or times out.

For a more detailed analysis, additional sub-scores are calculated. For each distribution type, the datasets are split into four groups according to their pathological fraction (P) and sample size (N):

1. Low P, low N:  $P \leq 0.2$ ,  $N \leq 5000$ .
2. Low P, high N:  $P \leq 0.2$ ,  $N > 5000$ .
3. High P, low N:  $P > 0.2$ ,  $N \leq 5000$ .
4. High P, high N:  $P > 0.2$ ,  $N > 5000$ .

Additionally, the accompanying R-package allows for calculating one score for each biomarker or individual scores for the different distribution types, sample sizes, or pathological fractions and overlaps. Furthermore, the computation time can be evaluated and compared on a pre-defined subset of test sets.

#### APPLICATION OF RIBENCH – COMPARISON OF 5 INDIRECT METHODS

To demonstrate the application of RIBench, we evaluated 5 different indirect methods. We focused on indirect methods in the group of primarily statistical approaches (6). All these methods share the basic assumption that the majority of test results are non-pathological and that the distribution of these can be modeled and identified despite an influence of pathological samples:

The Hoffmann method (22) was developed in 1963 as a probability paper method and assumes a normal distribution of the non-pathological test results. The modern methods, TML (8, 23–26), kosmic (10), TMC (9), and refineR (11) assume that the non-pathological test results can be modeled with a Box–Cox-transformed normal distribution (29). A short description as well as the parameters that were used for the different indirect methods are listed in Supplemental Methods and online Supplemental Table 1.

#### COMPARISON WITH A DIRECT METHOD

As a baseline, we provide a comparison to a (non-parametric) direct method with  $n = 120$  reference samples, which is considered the de facto standard for the minimum number of reference subjects in a direct RI study (5). The uncertainty of the direct method was estimated by Monte-Carlo simulation: 10 000 independent, random samples of size  $n = 120$  were drawn from the (theoretical) non-pathological distribution, characterized by the parameters mean ( $\mu$ ), standard deviation ( $\sigma$ ), lambda ( $\lambda$ ), and shift (S). For each sample, RLs were computed by determining the 2.5th and 97.5th (or 95th) percentiles using R's *quantile(type = 2)* function (30). These 10 000 nonparametric estimates were then processed in the exact same way as the estimates obtained from the indirect methods as described in section Evaluation and Performance Measure.

#### Results

Overall, the refineR algorithm achieved the best (lowest) benchmark score of 0.31, while the Hoffmann method, with applying the correct transformation, resulted in the highest z-score deviation of 1.18 (online Supplemental Table 2 and Fig. 3). The rate of implausible results, i.e., mean absolute z-score deviation  $>5$ , was highest for TML with 4.88% and lowest for refineR with 0.63%. Regarding the overall failure rate, TML could not find a result (or terminated, timed out, etc.) in 9.25% of all cases, while refineR always reported a result (Supplemental Table 2).

To investigate the influence of the proportion of pathological samples and the sample size, we split the data sets into 4 categories as defined above. Additionally, we analyzed each pathological fraction and sample size independently per distribution type as illustrated in Figs. 4 and 5 and Supplemental Figures 2 and 3. The results for each biomarker test set are depicted in online Supplemental Figures 4 and 5.

These results show that the performance of the indirect methods highly depends on the fraction of pathological data points (Figs. 3 and 4, and Supplemental Fig. 4). Up to a pathological fraction of 20%, all modern indirect methods achieved results comparable with the direct method. TMC, kosmic, and refineR also achieved comparable results up to a fraction of 30% or 40%, depending on the distribution type (Fig. 4). The Hoffmann method achieves comparable results up to a pathological fraction of about 5% when the true transformation parameter  $\lambda$  is specified. Using data with a higher pathological fraction (e.g., 50%) leads to increasing deviation from the truth for all indirect methods (Fig. 4 and Supplemental Fig. 4).

All modern indirect methods yield more precise results with larger sample size (Figs. 3 and 5 and

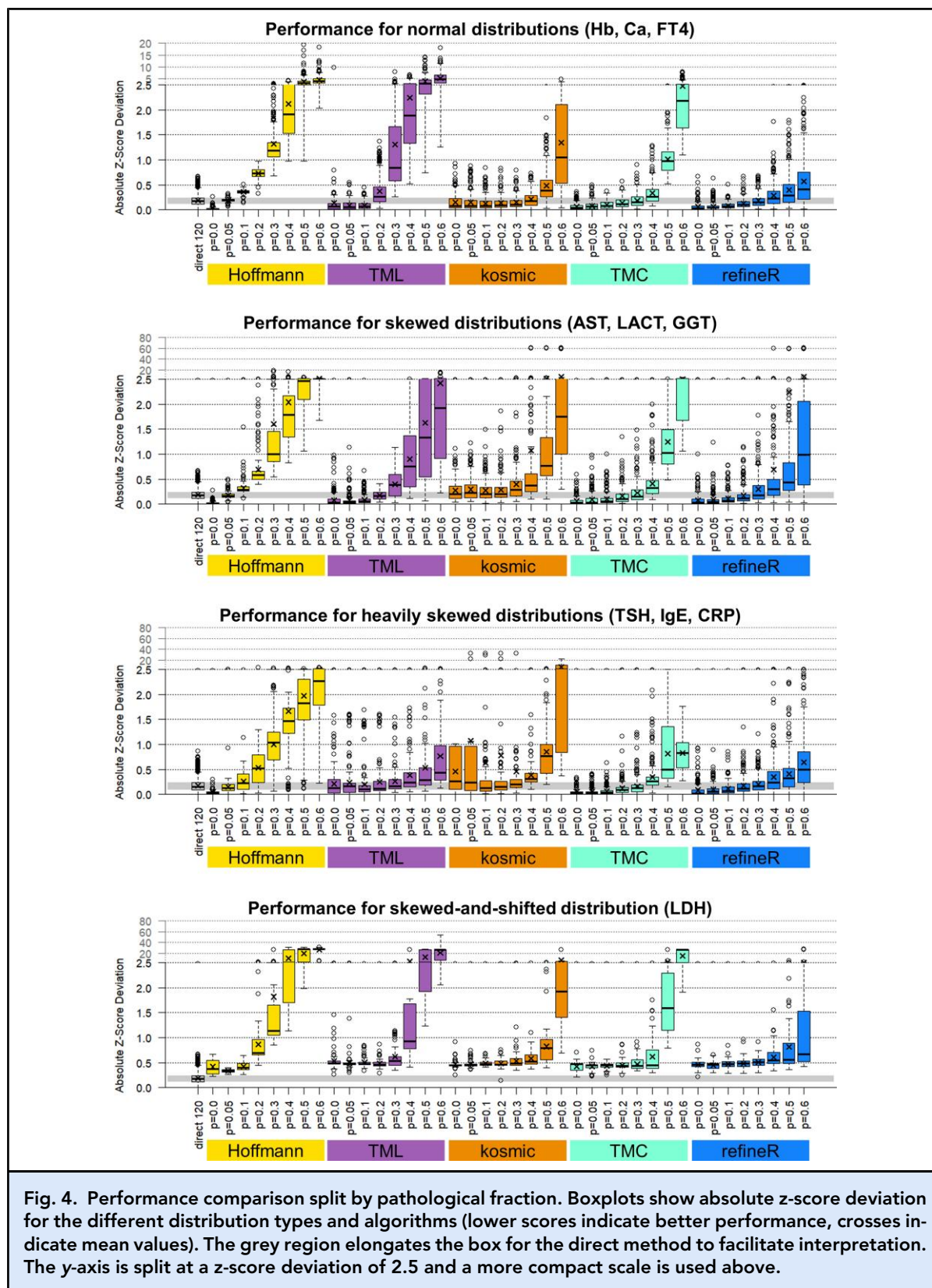


Fig. 4. Performance comparison split by pathological fraction. Boxplots show absolute z-score deviation for the different distribution types and algorithms (lower scores indicate better performance, crosses indicate mean values). The grey region elongates the box for the direct method to facilitate interpretation. The y-axis is split at a z-score deviation of 2.5 and a more compact scale is used above.

Supplemental Fig. 5). Using  $n \geq 50\,000$  data points leads to results superior to the direct method for most modern algorithms. Again, depending on distribution type and pathological fraction, the methods are capable of achieving precise results using fewer data points ( $n = 5000$ ), e.g., for normally distributed biomarkers or for test sets with a pathological fraction of  $\leq 20\%$ . In contrast, the Hoffmann method does not show improved results with an increased number of data points (Figs. 3 and 5 and Supplemental Fig. 5).

Overall, the influence of the pathological fraction on the performance is stronger than the influence of the sample size as shown by the difference in the benchmark score for low and high pathological fraction compared to the difference in a low and high sample size (Fig. 3 and Supplemental Table 2).

An in-depth analysis of the failure rate for the different distribution types shows that some algorithms are not robust for heavily skewed data sets. For these test sets, TML and kosmic fail or report an implausible result for 25.6% and 29.9% of test sets, respectively.

For the LDH test set, which is based on a skewed-and-shifted distribution (2-parameter Box–Cox transformation) for the non-pathological values, all indirect methods showed a systematic bias in the estimated RIs and did not achieve results comparable or superior to the direct method, independent of the amount of pathological and non-pathological data (Figs. 3 and 4). The refineR algorithm also provides an option for the 2-parameter Box–Cox transformation which yields results for LDH comparable to the direct method, while still providing robust results for the other distribution types (online Supplemental Table 3 and Supplemental Figs. 3–5).

To evaluate the computation time of the different indirect methods, a predefined subset comprising 500 test sets was utilized. The Hoffmann method achieved the lowest median computation time of 2.3 s per simulation, while TML resulted in the longest median runtime of 17.7 s (see online Supplemental Table 4).

## Discussion

Indirect methods represent an efficient, cost-effective, and easy-to-use alternative for laboratories to establish or verify RIs. Different indirect methods exist that are already widely implemented using RWD in different settings (12–16). However, there exists no comprehensive and objective evaluation using simulated data that can showcase the strengths and weaknesses of the different approaches and facilitate improvement of existing, or development of novel methods.

With the developed benchmarking suite Rlbench, we provide a tool that enables the systematic evaluation of

indirect methods covering a broad variety of test sets with varying difficulty. While Tan et al. (19) compared 8 methods for the exclusion or identification of the pathological distribution, and found the kosmic algorithm to perform best, we provide a more generic, comprehensive, and easily accessible tool to evaluate novel and existing methods.

Rlbench enables us to provide first answers to some important questions regarding the application of indirect methods in general. It reveals a systematic relationship between the performance of the different methods and the amount of input data/the fraction of pathological samples, with the latter having a greater influence. Up to a pathological fraction of 20%, the algorithms achieve results comparable or even superior to the direct method with  $n = 120$ . Having a pathological fraction  $\geq 50\%$  leads to large deviations from the truth for all algorithms (Figs. 3 and 4, Supplemental Tables 2 and 3, and Supplemental Fig. 4). These extreme cases violate the basic assumptions of the algorithms that the majority of samples originate from a non-pathological distribution, but help to highlight the limitations of the methods. In practice, additional meta-data (sample or patient information) may enable data pre-filtering to reduce the amount of pathological samples in the dataset (7).

The distribution type did not have a decisive influence on the overall performance, but rather the pathological fraction and overlap. Nevertheless, for heavily skewed distributions, TML and kosmic are less robust, possibly due to the internal use of a cumulative distribution function in the cost function (Supplemental Table 2 and Supplemental Fig. 3). Thus, for these distribution types, we recommend TMC or refineR. For skewed-and-shifted distributions, the refineR algorithm, offering the use of the 2-parameter Box–Cox transformation, showed the best performance. As the other evaluated indirect methods do not support this model and restrict the power parameter to values greater than or equal to zero, a systematic bias is introduced in the RI estimation for this distribution type. As stated by Ichihara et al. (31, 32) there exist probably more biomarkers like LDH in laboratory practice (i.e., biomarkers that follow a skewed distribution that is shifted away from zero).

Regarding a recommendation for sample size, no general answer can be provided, as the quality of results depends on the algorithm, the distribution type, and the pathological distribution(s). Nevertheless, in most scenarios, 5000 data points lead to robust results (Figs. 3 and 5, and Supplemental Fig. 5). In an optimal setting with few pathological data points ( $\leq 30\%$ ) and for biomarkers following an (approximately) normal distribution, kosmic, TMC, and refineR already produced results comparable to the direct method with  $n = 120$  with a sample size as low as  $n = 1000$  (Supplemental Fig. 2).



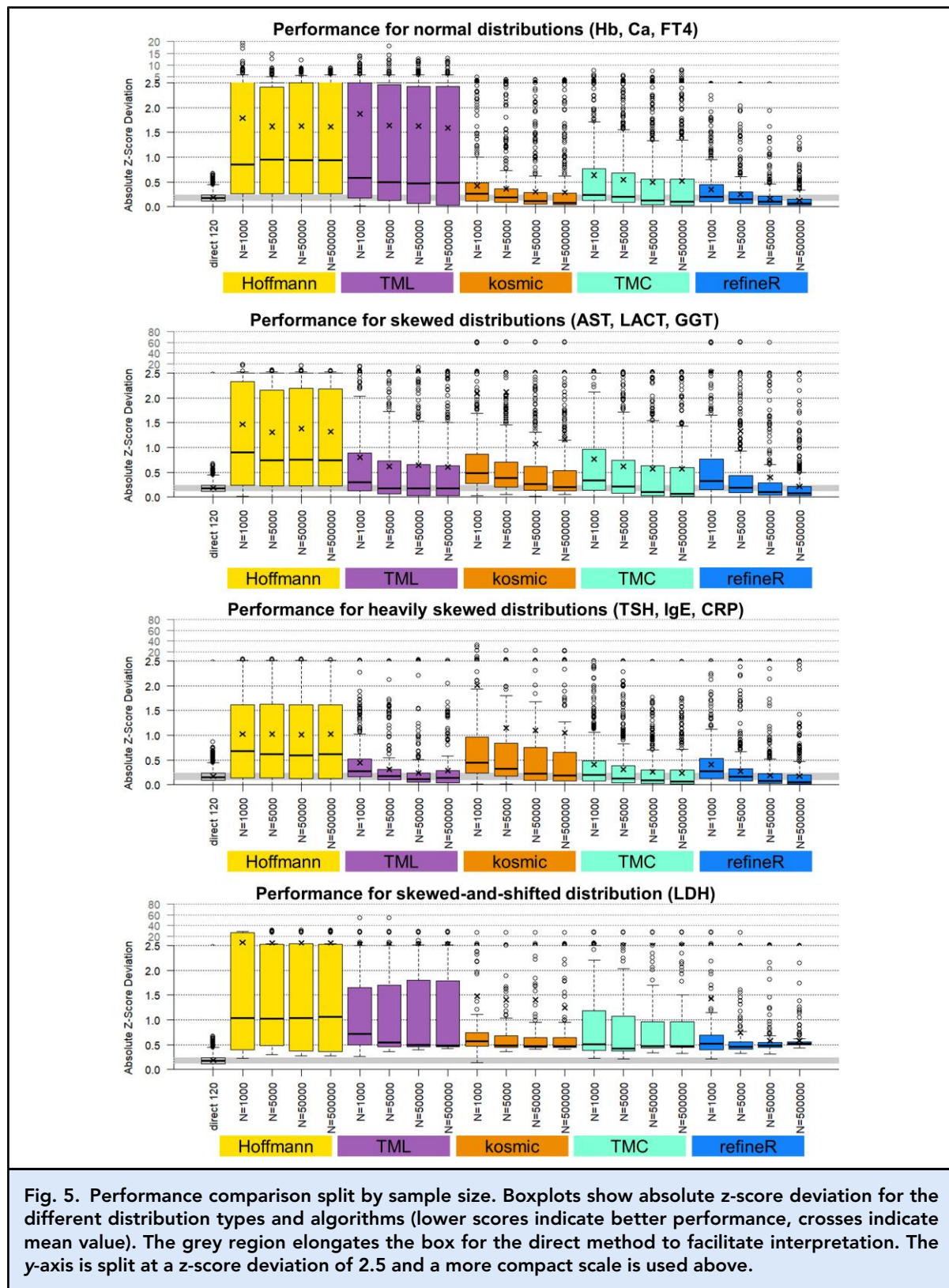


Fig. 5. Performance comparison split by sample size. Boxplots show absolute z-score deviation for the different distribution types and algorithms (lower scores indicate better performance, crosses indicate mean value). The grey region elongates the box for the direct method to facilitate interpretation. The y-axis is split at a z-score deviation of 2.5 and a more compact scale is used above.

While indirect methods can provide robust results in a wide variety of settings, individual challenging datasets can result in RI estimates with large deviations. We therefore highly recommend that results should be critically assessed by experts in laboratory medicine to detect such implausible results. In addition, it is important to emphasize that the overall benchmark score represents a high-level evaluation of an indirect method, and only the in-depth analysis including sub-scores provided by our R-package can unveil a comprehensive description of the individual strengths and weaknesses of an algorithm.

## LIMITATIONS

The provided benchmarking suite contains simulated data and thus enables a quantitative evaluation and comparison of the performance of indirect methods. However, simulations are not capable of representing all possible real-world scenarios. While our synthetic test sets provide nearly optimal input conditions, in most real-world settings a thorough data pre-filtering will be required to achieve best possible results, which is beyond the scope of our benchmarking suite and may not be feasible for every data set, as clinical data (e.g., diagnoses) may not always be available. Examples are appropriate physiological partitioning of age and sex and exclusion of multiple measurements of one subject (4). Moreover, the synthetic test sets are based on the assumption that the distribution of the non-pathological data can be modeled with a Box-Cox-transformed normal distribution. However, biomarkers may exist to which this assumption does not apply. Additionally, the pathological samples may be a composition of different distributions (i.e., be multimodal) due to the presence of multiple diseases in the input dataset. Thus, future work will focus on extending Rlbench to include additional distribution types, as well as multimodal pathological distributions.

We included 5 indirect methods explicitly designed for the problem of identifying the non-pathological distribution out of a mixed distribution of RWD. We did not cover more general approaches or outlier exclusion methods, but these can be included by the user employing the provided R-package.

For the Hoffmann method, no automated reference implementation has been available. Considering its reasonable complexity, we implemented the approach based on previous implementations (17), but did not apply a sophisticated optimization procedure. Additionally, we provided the correct transformation parameter to the Hoffmann method, which may bias results towards a “too good” performance, as this parameter is rarely known. However, even when providing this correct transformation parameter, the Hoffmann method led to results with the highest deviation. The Bhattacharya method (33) is not included, since no automated

reference implementation exists so far that could have been adapted for the purpose of this manuscript (34).

A further point of consideration is that the expected uncertainty from the direct method for a sample size of  $n = 120$  may be underestimated in our comparison, as we only included the statistical sampling error. In practice, challenges such as the ambiguous definition of “healthy,” and consequently the definition of inclusion and exclusion criteria, influence of the outlier exclusion approach (35), as well as dependency on covariates may increase uncertainty (4, 6).

## Conclusion

The presented benchmarking suite enables the standardized and systematic evaluation of existing and future indirect methods. Rlbench reveals the strengths and weaknesses of different methods, and answers important questions for the application of indirect methods, e.g., regarding sample size or pathological fraction. Hence, the provided R-package serves as a useful tool for future enhancements of existing and novel indirect methods, and ultimately as a valuable resource for improving RIs in laboratory practice.

## Supplemental Material

[Supplemental material](#) is available at *Clinical Chemistry* online.

**Nonstandard Abbreviations:** TML, truncated maximum likelihood; TMC, truncated minimum chi-square; RWD, real-world data; RI, reference interval; RL, reference limit.

**Author Contributions:** All authors confirmed they have contributed to the intellectual content of this paper and have met the following 4 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; (c) final approval of the published article; and (d) agreement to be accountable for all aspects of the article thus ensuring that questions related to the accuracy or integrity of any part of the article are appropriately investigated and resolved.

T. Ammer designed and implemented the benchmark suite, designed the simulated biomarker distributions, analyzed and interpreted the results, and wrote the manuscript. A. Schützenmeister, J. Zierk, C.M. Rank, and M. Rauh designed and supported the implementation of the benchmark suite, designed the simulations and interpreted the results. A. Schützenmeister, J. Zierk, C.M. Rank, M. Rauh, and H.-U. Prokosch supported the manuscript preparation and revision. All authors read and approved the final manuscript.

**Authors’ Disclosures or Potential Conflicts of Interest:** Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest: T. Ammer, A. Schützenmeister, H.-U. Prokosch, M. Rauh, C. M. Rank, and J. Zierk are the authors of the *refineR* algorithm; J. Zierk, M. Rauh, and H.-U. Prokosch are the main authors of the *kosmic* algorithm. All authors declare no further conflicts of interest.

**Employment or Leadership:** T. Ammer, A. Schützenmeister, and C.M. Rank are employees of Roche Diagnostics GmbH.

**Consultant or Advisory Role:** None declared.

**Stock Ownership:** C.M. Rank holds stocks/shares in F. Hoffmann-La Roche Ltd.

**Honoraria:** None declared.

**Research Funding:** The study was funded by Roche Diagnostic GmbH, Penzberg, Germany.

**Expert Testimony:** None declared.

**Patents:** None declared.

**Role of Sponsor:** No sponsor was declared.

**Acknowledgments:** We thank Thomas Helmbrecht (Roche Diagnostics GmbH, Penzberg, Germany) and Dusanka Kasapic (Roche Diagnostics International Ltd., Rotkreuz, Switzerland) for their valuable input, as well as Elizabeth A.S. Moser (Roche Diagnostics Operations, Indianapolis, IN, USA) for proofreading the manuscript.

## References

1. Tolan N V, Parnas ML, Baudhuin LM, Cervinski MA, Chan AS, Holmes DT, et al. "Big data" in laboratory medicine. *Clin Chem* 2015;61:1433–40.
2. Ma C, Wang X, Wu J, Cheng X, Xia L, Xue F, Qiu L. Real-world big-data studies in laboratory medicine: current status, application, and future considerations. *Clin Biochem* 2020;84:21–30.
3. Martinez-Sanchez L, Marques-Garcia F, Ozarda Y, Blanco A, Brouwer N, Canalias F, et al. Big data and reference intervals: rationale, current practices, harmonization and standardization prerequisites and future perspectives of indirect determination of reference intervals using routine data. *Adv Lab Med/Av en Med Lab* 2021. doi:10.1515/almed-2020-0034.
4. Jones GRD, Haeckel R, Loh TP, Sikaris K, Streichert T, Katayev A, et al. Indirect methods for reference interval determination—review and recommendations. *Clin Chem Lab Med* 2018;57:20–9.
5. CLSI. EP28-A3c Defining, establishing, and verifying reference intervals in the clinical laboratory: approved guideline-third edition. Wayne (PA): CLSI; 2010. <https://clsi.org/standards/products/method-evaluation/documents/ep28/>.
6. Zierk J, Metzler M, Rauh M. Data mining of pediatric reference intervals. *J Lab Med* 2021;45:311–7.
7. Haeckel R, Wosniok W, Streichert T, Members of the Section Guide Limits of the DGKL. Review of potentials and limitations of indirect approaches for estimating reference limits/intervals of quantitative procedures in laboratory medicine. *J Lab Med* 2021;45:35–53.
8. Arzideh F, Wosniok W, Haeckel R. Indirect reference intervals of plasma and serum thyrotropin (TSH) concentrations from intra-laboratory data bases from several German and Italian medical centres. *Clin Chem Lab Med* 2011;49:659–64.
9. Wosniok W, Haeckel R. A new indirect estimation of reference intervals: truncated minimum chi-square (TMC) approach. *Clin Chem Lab Med* 2019;57:1933–47.
10. Zierk J, Arzideh F, Kapsner LA, Prokosch HU, Metzler M, Rauh M. Reference interval estimation from mixed distributions using truncation points and the Kolmogorov-Smirnov distance (kosmic). *Sci Rep* 2020;10:1704.
11. Ammer T, Schützenmeister A, Prokosch H-U, Rauh M, Rank CM, Zierk J. refineR: A novel algorithm for reference interval estimation from real-world data. *Sci Rep* 2021;11:16023.
12. Shaw JLV, Cohen A, Konforte D, Binesh-Marvasti T, Colantonio DA, Adeli K. Validity of establishing pediatric reference intervals based on hospital patient data: a comparison of the modified Hoffmann approach to CALIPER reference intervals obtained in healthy children. *Clin Biochem* 2014;47:166–72.
13. Zierk J, Arzideh F, Rechenauer T, Haeckel R, Rascher W, Metzler M, et al. Age- and sex-specific dynamics in 22 hematologic and biochemical analytes from birth to adolescence. *Clin Chem* 2015;61:964–73.
14. Zierk J, Baum H, Bertram A, Boeker M, Buchwald A, Cario H, et al. High-resolution pediatric reference intervals for 15 biochemical analytes described using fractional polynomials. *Clin Chem Lab Med* 2021;59:1267–78.
15. Haeckel R, Wosniok W, Torge A, Junker R. Reference limits of high-sensitive cardiac troponin T indirectly estimated by a new approach applying data mining. A special example for measurands with a relatively high percentage of values at or below the detection limit. *J Lab Med* 2021;45: 87–94.
16. Moosmann J, Krusemark A, Dittrich S, Ammer T, Rauh M, Woelfle J, et al. Age- and sex-specific pediatric reference intervals for neutrophil-to-lymphocyte ratio, lymphocyte-to-monocyte ratio, and platelet-to-lymphocyte ratio. *Int J Lab Hematol* 2022;44:296–301.
17. Holmes DT, Buhr KA. Widespread incorrect implementation of the Hoffmann method, the correct approach, and modern alternatives. *Am J Clin Pathol* 2019; 151:328–36.
18. Ozarda Y, Ichihara K, Jones G, Streichert T, Ahmadian R, IFCC Committee on Reference Intervals and Decision Limits (C-RIDL). Comparison of reference intervals derived by direct and indirect methods based on compatible datasets obtained in Turkey. *Clin Chim Acta* 2021; 520:186–95.
19. Tan RZ, Markus C, Vasikaran S, Loh TP, APFCB Harmonization of Reference Intervals Working Group. Comparison of 8 methods for univariate statistical exclusion of pathological subpopulations for indirect reference intervals and biological variation studies. *Clin Biochem* 2022;103:16–24.
20. Kounev S, Lange K-D, von Kistowski J. Benchmarking basics. In: Kounev S, Lange K-D, von Kistowski J, editors. *Systems benchmarking: for scientists and engineers*. 1st Ed. Cham (Switzerland): Springer International Publishing; 2020. p. 3–21.
21. Haeckel R. Indirect approaches to estimate reference intervals. *J Lab Med* 2021;45: 31–3.
22. Hoffmann RG. Statistics in the practice of medicine. *JAMA* 1963;185:864–73.
23. Arzideh F, Wosniok W, Gurr E, Hinsch W, Schumann G, Weinstock N, Haeckel R. A plea for intra-laboratory reference limits. Part 2. A bimodal retrospective concept for determining reference limits from intra-laboratory databases demonstrated by catalytic activity concentrations of enzymes. *Clin Chem Lab Med* 2007;45:1043–57.
24. Arzideh F. Estimation of medical reference limits by truncated Gaussian and truncated power normal distributions [PhD thesis]. Bremen, Germany: Universität Bremen, 2008:171pp.
25. Arzideh F, Wosniok W, Haeckel R. Reference limits of plasma and serum creatinine concentrations from intra-laboratory data bases of several German and Italian medical centres: comparison between direct and indirect procedures. *Clin Chim Acta* 2010;411:215–21.
26. Arzideh F, Brandhorst G, Gurr E, Hinsch W, Hoff T, Roggenbuck L, et al. Ein verbesserter indirekter ansatz zur bestimmung von referenzgrenzen mittels intra-laboratorieller datensätze [An improved indirect approach for determining reference limits from intra-laboratory data bases exemplified by concentrations of electrolytes]. *J Lab Med* 2009;33:52–66.
27. Ceriotti F, Henny J, Queralto J, Ziyu S, Özarda Y, Chen B, et al. Common reference intervals for aspartate aminotransferase (AST), alanine aminotransferase (ALT) and  $\gamma$ -glutamyl transferase (GGT) in serum: results from an IFCC multicenter study. *Clin Chem Lab Med* 2010;48:1593–601.
28. Zetterstöm O, Johansson SGO. Ige concentrations measured by PRIST® in serum of healthy adults and in patients with respiratory allergy. A diagnostic approach. *Allergy* 1981;36:537–47.
29. Box GEP, Cox DR. An analysis of transformations. *J R Stat Soc Ser B* 1964;26: 211–52.
30. R Core Team. R: a language and environment for statistical computing. Vienna (Austria):



---

R Found. Stat. Comput; 2018. <https://cran.r-project.org/doc/FAQ/RFAQ.html#Citing-R> (Accessed August 2022).

31. Ichihara K, Boyd JC, IFCC Committee on Reference Intervals and Decision Limits (C-RIDL). An appraisal of statistical procedures used in derivation of reference intervals. *Clin Chem Lab Med* 2010;48:1537–51.
32. Ichihara K, Kawai T. Determination of reference intervals for 13 plasma proteins based on IFCC international reference preparation (CRM470) and NCCLS proposed guideline (C28-P, 1992): trial to select reference individuals by results of screening tests and application of maxim. *J Clin Lab Anal* 1996;10:110–7.
33. Bhattacharya CG. A simple method of resolution of a distribution into Gaussian components. *Biometrics* 1967;23:115–35.
34. Sikaris KA. Separating disease and health for indirect reference intervals. *J Lab Med* 2021;45:55–68.
35. Hickman PE, Koerbin G, Potter JM, Glasgow N, Cavanaugh JA, Abhayaratna WP, et al. Choice of statistical tools for outlier removal causes substantial changes in analyte reference intervals in healthy populations. *Clin Chem* 2020;66:1558–61.