Letter to the Editor

Data Mining Reference Intervals by ICD-10 Stratified Differential Distributions

To the Editor:

Personalized reference intervals are invaluable for the laboratory assessment of an increasingly aging population (1). To contextualize clinical laboratory results, accurate reference intervals (RIs) provide a baseline for comparison of the individual's test result. Demographic factors "age" and "sex" influence position and width of observed test result variance the most, yet the varying health status of patients also contributes to the residual variance of test results in stratified reference populations for indirect RI estimation and widens the inferred RIs. This is especially an issue for populations of elderly patients, where the prevalence of comorbidities increases with age (2). Determining what constitutes a "healthy" elderly person and what is a clinically insignificant or "normal" test result is a challenging task (3).

To address this problem, we propose use of the "Differential Distribution Method" (DDM). In addition to the patients' sex and age, it utilizes codes from the International Classification of Diseases coding system, Tenth Edition (ICD-10), along with patients' measurement results to differentiate the distribution of a "healthy" reference population from a general distribution. This is achieved by filtering out test results with statistically significant variance when grouped by co-occurrence of associated ICD-10 codes. These groups are formed based on any ICD-10 codes appearing among patients, whether as a single code or together with others. By partitioning values by groups of ICD-10 codes, these groups may provide an accurate representation of common comorbidities in a general clinical setting.

The DDM involves 4 stages: stratification, clustering, statistical testing, and RI inference. First, test results of a specified analyte are stratified by sex (male/female) and predefined age ranges (20 to 29, 30 to 39, … 70 to 79, and 80 to 89). For any combination of the factors "sex" and "age range" ("slice"), a total distribution covering all values ("Global Distribution", GD) is created and the entirety of unique diagnoses in ICD-10 code format are extracted (three-letter code). Second, the ICD-10 codes are clustered using the word2vec natural language processing (NLP) algorithm based on the resulting similarity matrix. Third, for each individual diagnosis in the slice, test results are grouped by the obtained clusters and compared to the GD, assessing whether the distributions differ significantly. Fourth, test results showing statistical significance, as established by *t*-testing with P values <0.05 (adjusted for multiple comparisons), are removed from the individual age- and sex-stratified slices to create the "Differential Distributions" (DD), from which RIs are estimated by an iterative parametric approach (4).

The data set used for illustration consisted of 226 527 test results of plasma creatinine from inpatients, 20 to 90 years of age, described elsewhere (5). Overall variance increased with age for creatinine levels of both female and male study participants. Among the ICD-10 codes that were associated with significantly elevated test results relative to the GD, codes referred to renal (N17, N18, and N19) and systemic diseases (E66, I10, and I50) as well as traumatic injuries or therapeutic intervention (S01, S02, S06, V99, Y84, and Z98). The creation of DDs resulted in removing large portions of results, which slightly increased the 90% confidence intervals; however, the inferred RIs were notably narrower across all strata (Table 1).

Routine testing enables the identification of "pathological" values indicative of an underlying health condition but also generates a substantial volume of "non-pathological" values. The initial creatinine data showed a concentration of values within the expected "healthy" range yet presented high variance and long distribution tails. Comorbidity-associated removal of results that originate from a significantly different distribution reduced the variance in the stratified reference populations compared to the GD. The DDM provides valuable insights into which ICD-10 subpopulations skew the GD the most and generate distribution tails, i.e., which ICD-10 diagnoses likely contribute mostly pathological values.

The clustering reliably produces ICD-10 code groups, which may have previously unconsidered implications on their own regarding co-occurrence of particular diagnoses. Further research is necessary to assess the interpretation of the clustering, as the hierarchical approach of the DDM does not incorporate the clinical context from the start. Additionally, the qualitative aspects of ICD-10 codes used in this retrospective study require further examination, particularly whether these are consistently recorded, accurately reflect the patient's conditions at the time of testing, or if they were only subsequently assigned based on the test results. Tailoring methods at this level of detail is the topic of future research. Limitations of the use of ICD-10 codes are inherent in the described procedures too. The DDM tailors RIs specifically

Table 1. Creatinine reference intervals ($X_{2.5th}$ and $X_{97.5th}$ percentiles in mg/dL) for indicated age ranges inferred from the various distributions.[a]

	Global Distribution (GD)			Differential Distribution (DD) with clustering		
Age range (in years)	n	X2.5th	X97.5th	n	X2.5th	X97.5th
Female						
20-29	8858	0.33 (0.32-0.33)	0.93 (0.92-0.93)	837	0.43 (0.42-0.44)	0.87 (0.86-0.88)
30-39	13 505	0.33 (0.33-0.34)	0.93 (0.92-0.93)	888	0.42 (0.40-0.43)	0.94 (0.92-0.95)
40-49	10 379	0.41 (0.40-0.41)	0.97 (0.97-0.98)	4164	0.45 (0.45-0.46)	0.97 (0.96-0.98)
50-59	14 862	0.42 (0.42-0.43)	1.00 (0.99-0.99)	3349	0.45 (0.45-0.46)	0.98 (0.98-0.99)
60-69	18 037	0.41 (0.40-0.41)	1.05 (1.05-1.06)	1481	0.45 (0.45-0.47)	1.02 (1.01-1.04)
70-79	22 113	0.41 (0.40-0.41)	1.13 (1.13-1.14)	657	0.45 (0.44-0.48)	1.09 (1.07-1.11)
80-89	17 240	0.40 (0.39-0.40)	1.27 (1.26-1.28)	313	0.44 (0.41-0.48)	1.17 (1.13-1.20)
Male						
20-29	5827	0.57 (0.56-0.58)	1.17 (1.16-1.17)	2638	0.62 (0.62-0.63)	1.14 (1.13-1.15)
30-39	7434	0.55 (0.55-0.56)	1.19 (1.18-1.20)	3354	0.59 (0.59-0.60)	1.18 (1.17-1.19)
40-49	11 854	0.55 (0.55-0.56)	1.19 (1.19-1.20)	3489	0.59 (0.58-0.60)	1.19 (1.18-1.19)
50-59	23 049	0.52 (0.51-0.52)	1.24 (1.24-1.25)	2402	0.58 (0.56-0.58)	1.21 (1.20-1.22)
60-69	31 055	0.51 (0.50-0.51)	1.30 (1.30-1.31)	711	0.60 (0.58-0.62)	1.22 (1.20-1.24)
70-79	31 754	0.53 (0.52-0.53)	1.39 (1.39-1.40)	708	0.58 (0.55-0.60)	1.35 (1.32-1.37)
80-89	15 743	0.54 (0.53-0.54)	1.55 (1.54-1.55)	641	0.54 (0.52-0.58)	1.40 (1.37-1.43)

[a]Reference interval estimates are drawn from the Global Distribution (GD, left) and the Differential Distribution (DD) generated using hierarchical clustering with 800 clusters (right). Obtained RIs also contain the 90% confidence intervals in brackets, indicating their associated precision. Measurements were converted with 1 μmol/L = 0.01131 mg/dL.

for the locally admitted patient population, by accounting for age-related and circumstantial physiological changes, while excluding severe pathologies (diabetes, hypertension, traumatic injuries, etc.). By integrating ICD-10 codes, this enhances diagnostic utility and reduces the need for further testing by providing more personalized comparison tools. It is essential to consider the patients' health status not only during the diagnostic process but likewise when establishing more personalized RIs: The better the RIs fit the comorbidity patterns, the more useful they will be for clinical practice.

Ethical approval

This study received an ethics waiver from the cantonal ethics committee of Bern (Business Administration System for Ethics Committees; BASEC-Nr: Req-2020-00630).

Nonstandard Abbreviations: RI, reference interval; DDM, Differential Distribution Method; ICD-10, system International Classification of Diseases coding system, Tenth Edition; GD, Global Distribution.

Author Contributions: *The corresponding author takes full responsibility that all authors on this publication have met the following required criteria of eligibility for authorship: (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. Nobody who qualifies for authorship has been omitted from the list.*

Tobias Blatter (Data curation-Equal, Formal analysis-Equal, Investigation-Supporting, Methodology-Equal, Software-Equal, Validation-Equal, Visualization-Equal, Writing—original draft-Lead, Writing—review & editing-Lead), David Schär (Formal analysis-Equal, Investigation-Equal, Methodology-Equal, Software-Equal, Validation-Equal, Visualization-Equal, Writing—original draft-Equal, Writing—review & editing-Equal), Harald Witte (Data curation-Lead, Project administration-Lead, Writing—original draft-Equal, Writing—review & editing-Equal), Christos Nakas (Conceptualization-Equal, Formal analysis-Supporting, Methodology-Equal, Supervision-Equal, Writing—original draft-Equal, Writing—review & editing-Equal), and Alexander Leichtle (Conceptualization-Equal, Funding acquisition-Lead, Methodology-Equal, Supervision-Equal, Writing—review & editing-Equal)

Authors' Disclosures or Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the author disclosure form.*

Research Funding: The research project was funded by the Swiss Personalized Health Network (SPHN), grant number 2018DEV22 and the University Hospital Bern.

Disclosures: A.B. Leichtle has received funding for other projects from the Bern Center of Precision Medicine, SPHN, the University Bern Interdisciplinary Grant, the Bern Multidisciplinary Center of Infectious Diseases (MCID), and the Swiss National Science Foundation, and is a member of the Task Force on Global Reference Interval Database (TF-GRID) of the International Federation of Clinical Chemistry (IFCC).

Acknowledgments: Collaborators: Prof. Dr. Jivko Stoyanov, Swiss Paraplegic Research, Nottwil, Switzerland; Prof. Dr. Martin Hersberger, University Children's Hospital Zurich, Switzerland, Nicolas Rosat, Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland.

The authors would like to thank the SPHN Data Coordination Center and the team of the BioMedIT node at ETH Zürich for support and, most importantly, all patients for sharing their data.

References

1. Timbrell NE. The role and limitations of the reference interval within clinical chemistry and its reliability for disease detection. Br J Biomed Sci 2024;81:12339.
2. Ioakeim-Skoufa I, Clerencia-Sierra M, Moreno-Juste A, Elías de Molins Peña C, Poblador-Plou B, Aza-Pascual-Salcedo M, et al. Multimorbidity clusters in the oldest old: results from the EpiChron cohort. Int J Environ Res Public Health 2022;19:10180.
3. Jørgensen LGM, Brandslund I, Hyltoft Petersen P. Should we maintain the 95 percent reference intervals in the era of wellness testing? A concept paper. Clin Chem Lab Med 2004;42:747–51.
4. Ichihara K, Kawai T. An iterative method for improved estimation of the mean of peer-group distributions in proficiency testing. Clin Chem Lab Med 2005;43: 412–21.
5. Blatter TU, Witte H, Fasquelle-Lopez J, Raisaro JL, Leichtle AB. The BioRef infrastructure, a framework for real-time, federated, privacy-preserving, and personalized reference intervals: design, development, and application. J Med Internet Res 2023; 25:e47254.

Tobias U. Blatter[a,b,*], **David Schär**[a], **Harald Witte**[a], **Christos T. Nakas**[a,c], **and Alexander B. Leichtle**[a,d]

[a]University Institute of Clinical Chemistry, Inselspital Bern, University Hospital and University of Bern, Bern, Switzerland
[b]Graduate School for Health Sciences (GHS), University of Bern, Bern, Switzerland
[c]Laboratory of Biometry, University of Thessaly, Volos, Greece
[d]Center for Artificial Intelligence in Medicine (CAIM), University of Bern, Bern, Switzerland

[*]Address correspondence to this author at: University Institute of Clinical Chemistry, Inselspital Bern, University Hospital and University of Bern, Freiburgstrasse, Bern 3010, Switzerland. Tel +41-31-632-83-30; e-mail tobias.blatter@extern.insel.ch.

https://doi.org/10.1093/clinchem/hvae089