TITLE PAGE

Running title

Diagnostic test accuracy in longitudinal study settings: Theoretical approaches with use cases from clinical practice

Full name, department, institution, city, and country of all co-authors

Julia Böhnke[1], Antonia Zapf[2], Katharina Kramer[3], Philipp Weber[2], ELISE Study Group[4], André Karch[1*], Nicole Rübsamen[1*]

1. Institute of Epidemiology and Social Medicine, University of Münster, Germany, Albert-Schweitzer-Campus 1, 48149 Münster, Germany
2. Department of Medical Biometry and Epidemiology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
3. Mathematical Statistics and Artificial Intelligence in Medicine, University of Augsburg, Augsburg, Germany
4. ELISE study group members

   Louisa Bode[a]; Marcel Mast[a]; Antje Wulff[a,f]; Michael Marschollek[a]; Sven Schamer[b]; Henning Rathert[b]; Thomas Jack[b]; Philipp Beerbaum[b]; Nicole Rübsamen[c]; Julia Böhnke[c]; André Karch[c]; Pronaya Prosun Das[d]; Lena Wiese[d]; Christian Groszweski-Anders[e]; Andreas Haller[e]; Torsten Frank[e]

[a] *Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Hannover, Germany*

[b] *Department of Pediatric Cardiology and Intensive Care Medicine, Hannover Medical School, Hannover Germany*

[c] *Institute of Epidemiology and Social Medicine, University of Münster, Münster, Germany*

[d] *Research Group Bioinformatics, Fraunhofer Institute for Toxicology and Experimental Medicine, Hannover, Germany*

[e] *medisite GmbH, Hannover, Germany*

[f] *Big Data in Medicine, Department of Health Services Research, School of Medicine and Health Sciences, Carl von Ossietzky University Oldenburg, Oldenburg, Germany*

* *These authors contributed equally to this work.*

ORCiD of all authors

Author	ORCiD
Julia Böhnke	https://orcid.org/0000-0003-1249-4581
Antonia Zapf	https://orcid.org/0000-0002-8467-0508
Katharina Kramer	https://orcid.org/0000-0002-9734-0065
Philipp Weber	https://orcid.org/0000-0002-9520-6284
André Karch	https://orcid.org/0000-0003-3014-8543
Nicole Rübsamen	https://orcid.org/0000-0003-2198-5577
Louisa Bode	https://orcid.org/0000-0002-4570-9588
Marcel Mast	https://orcid.org/0000-0001-6478-2230
Antje Wulff	https://orcid.org/0000-0002-2550-2627
Michael Marschollek	https://orcid.org/0000-0002-5921-3073
Sven Schamer	https://orcid.org/0000-0003-3626-561X
Henning Rathert	https.//orcid.org/0000-0001-9610-0257
Thomas Jack	https://orcid.org/0000-0003-0576-327X
Philipp Beerbaum	https://orcid.org/0000-0001-9090-0392
Pronaya Prosun Das	https://orcid.org/0000-0003-0165-5167
Lena Wiese	https://orcid.org/0000-0003-3515-9209
Christian Groszweski-Anders	---
Andreas Haller	---
Torsten Frank	---

35 **Full name, postal address, e-mail and telephone number of the corresponding authors**

36 Julia Böhnke, MSc

37 Institute of Epidemiology and Social Medicine

38 University of Münster

39 Albert-Schweitzer-Campus 1

40 48149 Münster

41 Germany

42 Phone: +49 251 83 57033

43 Fax: +49 251 83 55300

44 Mail: boehnkej@uni-muenster.de

45 ORCID: 0000-0003-1249-4581

46

47 **Word count**

48 199 of 200 words in Abstract

49 3,942 of 4,000 words in main article (excludes the title page, abstract, table(s), acknowledgments,

50 contributions and references)

51

52 **Article type**

53 Practice of Epidemiology and Methodology

ABSTRACT

In this study we evaluate how to estimate diagnostic test accuracy (DTA) correctly in the presence of longitudinal patient data (i.e., repeated test applications per patient). We used a nonparametric approach to estimate sensitivity and specificity of diagnostic tests for three use cases with different characteristics (i.e., episode length and intervals between episodes): 1) systemic inflammatory response syndrome, 2) depression, and 3) epilepsy. DTA was estimated on the levels '*time*', '*event*', and '*patient-time*' for each diagnosis, representing different research questions. A comparison of DTA for these levels per and across use cases showed variations in the estimates, which resulted from the used level, the time unit (i.e., per minute/hour/day), the resulting number of observations per patient, and the diagnosis-specific characteristics. Researchers need to predefine their choices (i.e., estimation levels and time units) based on their individual research aims, including the estimand definitions, and give an appropriate rationale considering the diagnosis-specific characteristics of the target outcomes and the number of observations per patient to make sure that unbiased and clinically relevant measures are communicated. Nonetheless, researchers could report the DTA of the test using more than one estimation level and/or time unit if this still complies with the research aim.

KEYWORDS

Diagnostic study, diagnostic test accuracy, longitudinal study, data cluster, nonparametric method, estimation level

1. INTRODUCTION

A diagnostic test (DT) can be any device (e.g., biomarker quantification of bodily fluids, magnetic resonance imaging, or clinical decision support system [CDSS])[1–3] with which healthcare professionals can classify a target condition (e.g., diseased vs disease-free)[1–6] and make an informed decision based on the test's result. Each test requires to be assessed for its diagnostic accuracy (i.e., sensitivity and specificity) before its usage in daily practice within medical settings[7]. Ideally, a DT should provide a correct classification of the target condition (i.e., *true positives* [TP], *true negatives* [TN], Appendix 1) while being safe and effective in its diagnostic performance[2,3,5]; thus, the quantity of *false positive* (FP) and *false negative* (FN) test results should be minimal[5]. Misdiagnoses can have serious consequences for the patient's health[2,5], including mental distress[7], and/or for a country's healthcare system (e.g., unnecessary costs)[2].

The diagnostic validity of the DT (referred to as *index test*, IT) is best assessed in a diagnostic test accuracy (DTA) study using an established, carefully selected *reference standard* (RS) as the ground truth[5,7]. To minimize potential influences, both tests should be blinded to each other, and performed without time delay to avoid diagnostic differences caused by temporal changes in the target condition[2,5]. Conducting the evaluation of a DT with a DTA study provides healthcare professionals with the necessary information on the DT's performance so that they can make an informed decision[5]. Information on test performance is usually reported in terms of sensitivity and specificity (Appendix 1 for key terminology of DTA studies).

Lately, researchers showed that many DTA studies are of low quality, do not necessarily represent the clinical situation of interest, and/or are associated with a considerable risk of bias[8,9]. As a consequence, the DT under review might not be used in practice, or the research may be

involuntarily distorted and most likely overoptimistic about the IT's performance[9,10]. Particularly, repeated measurements per patient (see Appendix 2 for examples) require adequate DTA assessment as the within-person correlation can inflate the DT's uncorrected accuracy compared to only including a single measurement per patient in the DTA estimation[11–13]. A systematic review on studies evaluating the DTA of CDSS highlighted that most DTA studies did not report sufficient information on the usage of or adjustment for longitudinal data (i.e., repeated measurements per patients with disease-free and/or diseased periods) in the DTA estimation[8]. The DTA studies that accounted for longitudinal data used various methods to adjust their DTA estimates[14].

Treatment effects must also be considered: An early intervention may hinder the onset of the target condition while treatment after diagnosis may cause a health improvement so that the diagnostic status may change from diseased to disease-free. An a priori definition of the estimand that is the target of estimation to address the scientific question of interest posed by the study objective[15–18] is, therefore, necessary.

This study systematically evaluates how to analyze and report longitudinal data from DTA studies using datasets on systematic inflammatory response syndrome (SIRS), depression, and epilepsy as use cases. The longitudinal data challenge will be addressed by:

- presenting three DTA estimation levels (i.e., time-level, event-level, and patient-time-level) and their respective DTA estimations, and
- introducing a nonparametric method based on research by Konietschke & Brunner[19] and Lange[11].

Note that other methods of DTA estimation accounting for longitudinal data[10] are available (not addressed in this paper); regardless, the approaches of this article (choice of time unit and estimation level) apply.

2. METHODS

This study is reported in accordance with the items of the 2015 Standard for Reporting Diagnostic Accuracy guideline[20] (Appendix 3). In the following, we use the following nomenclature: A "time unit" is chosen by the researcher, i.e. the diagnostic status is assessed every minute/hour/day etc. A "time point" refers to a specific minute/hour/day within the longitudinal setting, e.g. hour 24 of a patient's stay. A "block" is an aggregation of labeled time points based on the rules explained below.

2.1 DTA estimation levels

We present three DTA estimation levels (Figure 1 and Appendix 4) determining an IT's performance using longitudinal patient data.

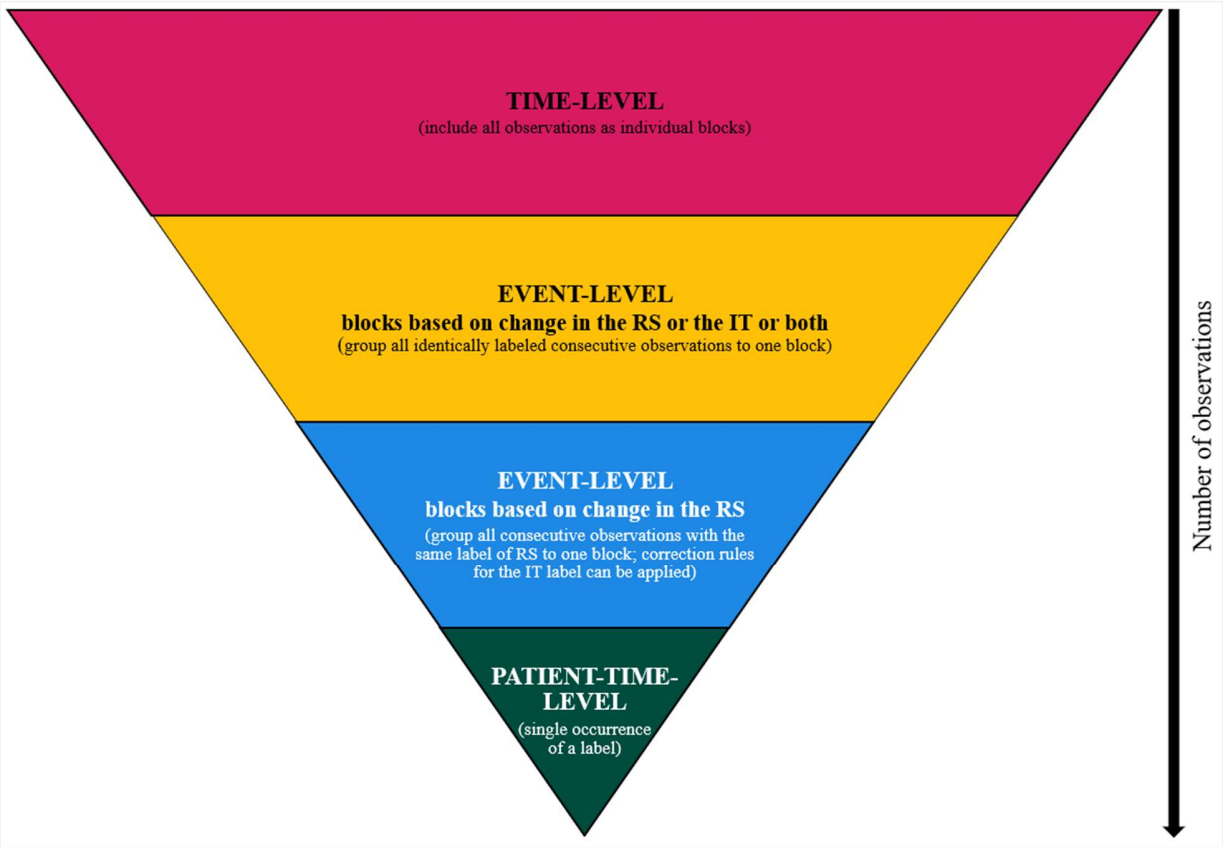


Figure 1: Visualization of the data structure and its subunits that are included in the diagnostic test accuracy estimation. Two options for the event-level are presented that differ regarding their groupings of labeled time points into blocks.

2.1.1 Time-level

The time-level provides labels (i.e., TP, TN, FP, or FN) for every time point. This level's estimand is the diagnostic status per time point without any aggregation.

2.1.2 Event-level

The event-level aggregates consecutive, labeled time points based on diagnostic status; thus, per patient, the minimum block length is one time unit and the maximum block length is equivalent to

the total of all time points (i.e., no change in diagnostic status). This level requires that the estimand is a change in the diagnostic status.

In the following, we differentiate between blocks based on the RS alone, or on both, the IT and the RS.

2.1.2.1 Blocks based on RS

The time point where the RS changes its diagnostic status determines the end of the previous block and the start of the new block. With this definition, the result of the RS is assumed to be known while the result of the IT is a random variable that follows a Bernoulli distribution.

For DTA estimation, the time point labels per block are summarized into one single label. FP and FN labels always overrule TP and TN labels as they indicate differences between the tests. The summarized block labels are used in the DTA estimation. This labeling penalizes any differences (i.e., early/late episode start/end, etc.) between the DTs by having an increase of FP and FN labels. We can control for this by applying modifying rules, for example with applying a clinician-based tolerance margin rule so that if the IT starts/ends within the tolerance margin of the RS, the IT's diagnostic status of the specific time points is changed in accordance with the RS's diagnostic status (i.e., no "punishment" if IT starts/ends too early/late). However, if the IT starts/ends outside of the tolerance margin, the IT's diagnostic status of these specific time points remains unchanged. A %-correctness rule can also be applied according to which the IT's diagnostic status per patient is corrected in accordance with the RS's diagnostic status if at minimum $P_{diseased}$ percent of single time points per a diseased block and at minimum $P_{disease\text{-}free}$ percent of single time points per a disease-free block are correctly classified. The P's are diagnosis-specific. For our analysis, we used a ±1 time interval tolerance margin around the RS's episode start and end and an 85% correction rule for diseased and disease-free blocks (see Appendix 5 for analyses using other modifying rules). Afterwards, the labels are summarized into a single label and used for the DTA estimation.

2.1.2.2 Blocks based on IT and RS

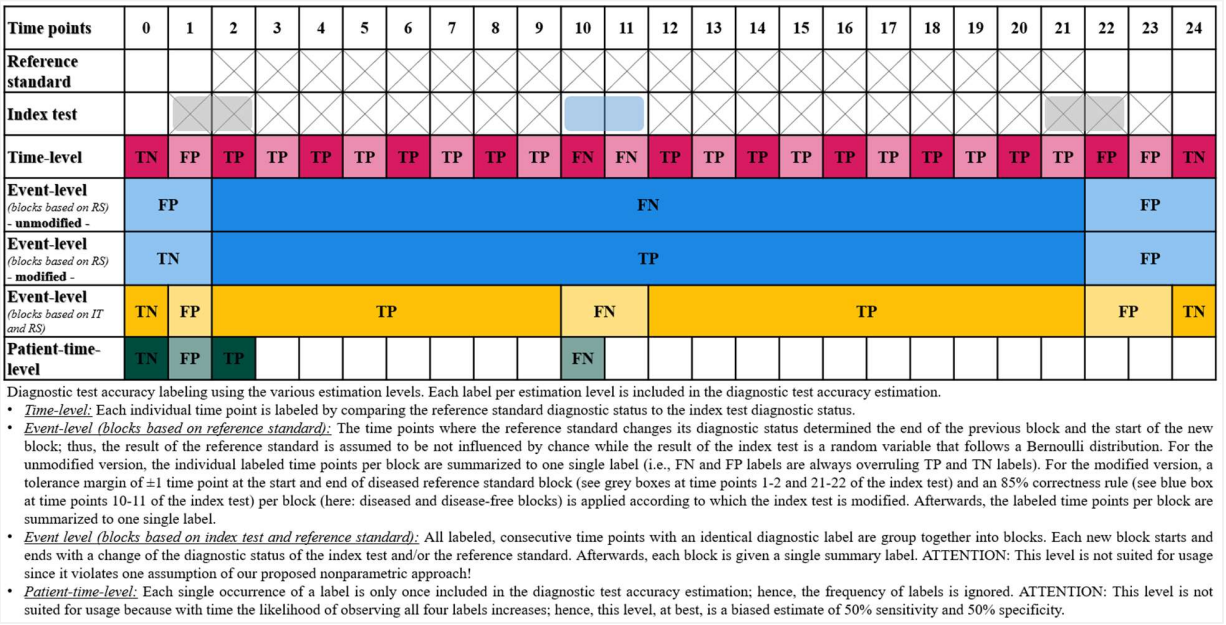
At first, each time point is labeled and then all consecutive time points with an identical label are grouped together into blocks. Each new block starts and ends with a change of the diagnostic status of the IT and/or the RS. Afterwards, each block is given a single summary label that is used for the DTA estimation. Modifying rules can be applied. With this definition, both the result of the RS and the results of the IT are random variables, which violates one assumption of our proposed nonparametric approach.

2.1.3 Patient-time-level

The patient-time-level summarizes the occurrence of all labels per patient during the defined period; thus, a patient adds at minimum one label (i.e., either TP, FP, FN, or TN) or at maximum all four labels once to the DTA estimation. This level's estimand is the occurrence of the different possible labels, without considering their respective frequency. It is not suited for usage because with time the probability of observing all four labels increases; hence, this level, at best, is a biased estimate of 50% sensitivity and 50% specificity.

2.1.4 Example of labeling per estimation level

An example of labeling of the three levels is displayed in Figure 2. Each crossed out cell marks the presence of the target condition according to the respective test at that particular time point. Below are the labeled units per level, as described previously, which are used for the DTA estimation. The event-level with blocks based on IT and RS as well as the patient-time-level are shown for illustration only; they should not be used in clinical practice.



Diagnostic test accuracy labeling using the various estimation levels. Each label per estimation level is included in the diagnostic test accuracy estimation.

- *Time-level:* Each individual time point is labeled by comparing the reference standard diagnostic status to the index test diagnostic status.
- *Event-level (blocks based on reference standard):* The time points where the reference standard changes its diagnostic status determined the end of the previous block and the start of the new block; thus, the result of the reference standard is assumed to be not influenced by chance while the result of the index test is a random variable that follows a Bernoulli distribution. For the unmodified version, the individual labeled time points per block are summarized to one single label (i.e., FN and FP labels are always overruling TP and TN labels). For the modified version, a tolerance margin of ±1 time point at the start and end of diseased reference standard block (see grey boxes at time points 1-2 and 21-22 of the index test) and an 85% correctness rule (see blue box at time points 10-11 of the index test) per block (here: diseased and disease-free blocks) is applied according to which the index test is modified. Afterwards, the labeled time points per block are summarized to one single label.
- *Event level (blocks based on index test and reference standard):* All labeled, consecutive time points with an identical diagnostic label are group together into blocks. Each new block starts and ends with a change of the diagnostic status of the index test and/or the reference standard. Afterwards, each block is given a single summary label. ATTENTION: This level is not suited for usage since it violates one assumption of our proposed nonparametric approach!
- *Patient-time-level:* Each single occurrence of a label is only once included in the diagnostic test accuracy estimation; hence, the frequency of labels is ignored. ATTENTION: This level is not suited for usage because with time the likelihood of observing all four labels increases; hence, this level, at best, is a biased estimate of 50% sensitivity and 50% specificity.

Figure 2: Example of labeling on the three levels. The time-level adds 18 true positive (TP), 3 false positive (FP), 2 false negative (FN), and 2 true negative (TN) observations to the DTA estimation. The DTA estimation of the event-level using blocks based on RS adds 1 TP, 1 FP, and 1 TN to the DTA estimation while the event level using blocks based on both tests adds 2 TP, 2 FP, 1 FN, and 2 TN observations. On the patient-time-level, all four labels were observed; thus, this patient adds one observation to each label.

2.2 The nonparametric approach

The DTA can be estimated using the nonparametric approach based on research by Konietschke & Brunner[19] and Lange[11] which is robust and reliable even when accounting for intra- and interclass correlations[21]. Konietschke & Brunner[19,21] proposed a categorization of participants into three cluster groups, regardless of the individual participant's number of repeated measurements:

- 'Absent' (ic_0): incomplete cases with target condition consistently absent (i.e., patient was consistently disease-free during the total observation period; these cases are "incomplete" because diseased phases are missing).
- 'Present' (ic_1): incomplete cases with target condition consistently present (i.e., patient was consistently diseased during the total observation period; these cases are "incomplete" because disease-free phases are missing).
- 'Mix' (c): complete cases with target condition both present and absent (i.e., the patient experienced diseased and disease-free phases during the total observation period).

In the DTA estimation, this method uses a unified nonparametric model to estimate the area under the curve, sensitivity, and specificity accounting for a longitudinal data format[11,22]. This approach applies a nonparametric rank statistic while accounting for the clustering by using the weighted estimation strategy (i.e., weighting by size of the clusters; thus, larger clusters have larger weights than smaller ones)[11,21]. This allows assigning an equal weight to all subunits of the same cluster[21]. Each DTA estimate is presented with its 95% logit Wald confidence interval (CI). For details on the method, we refer to [11,19]. The R code for the analyses is provided at https://zivgitlab.uni-muenster.de/ruebsame/dta_longitudinal_data_methods.

2.3 The datasets

We used three publicly available datasets to show the application of our proposed methods across different medical fields. Dataset descriptions, dataset labeling, and information on ITs and RSs are presented in Appendix 5.

2.3.1 SIRS dataset

The SIRS dataset includes 168 male and female pediatric patients (0-17 years). All participants were consecutively recruited at a single study center (i.e., Department of Pediatric Cardiology and Pediatric Intensive Care at Hannover Medical School, Germany) between 2018-08-01 and 2019-03-31. A total of 101 of the 168 patients developed at least one SIRS episode at any time during their inpatient stay at the study center. For details, we refer to [23–25]. We used the data of 36 patients (26 diseased individuals) to ensure comparability with the other datasets regarding the sample size.

2.3.2 Depression dataset

The depression dataset includes records of 55 adult patients of which 23 experienced a depressive episode (5 inpatients and 18 outpatients) and 32 individuals did not (23 hospital employees, 5 students, and 4 former, currently non-depressive, patients). All individuals were recruited while being treated at the Department of Psychiatry of the Haukeland University, Norway. The depressive patients were equipped with a wearable sensor that recorded the patients' motor activity per minute since depressive people tend to decrease their personal activity (i.e., reduced active during day-time hours). In total, activity data from 693 days were recorded (diseased: 291 days; controls: 402 days)[26]. For this study's purpose, the dataset included all cases and only the first 10 controls (33 patients in total).

2.3.3 Epilepsy dataset

The epilepsy dataset entails electroencephalogram (EEG) recordings of 22 pediatric and young adult patients (5 males, 3-22 years; 17 females, 1.5-19 years; ID chb01 and chb21 are from the same person) with intractable seizures of the Boston Children's Hospital, USA. One extra patient was added later. Each patient was likely to develop an epileptic episode due to having stopped the

anti-seizure medication under medical supervision in an inpatient setting. A total of 197 episodes were recorded (i.e., each patient had ≥2 episodes). EEG-signals were recorded at a frequency of 256 samples per second with 16-bit resolution[27]. Modifications were applied to the dataset to meet this study's research purpose: Six additional disease-free synthetic patient records were added to have a sample size of 30 patients.

2.4 Analysis

The assessment of the datasets by the ITs and RSs (Appendix 4 and 5) was consistently applied to the (modified) datasets. Missing values of the IT's and RS's assessments were not observed. Indeterminate test results were not registered.

Sensitivities and specificities were estimated for each diagnosis per time unit (i.e., minute, hour, and day) and per estimation level (i.e., time-level, event-level, and patient-time-level) using the nonparametric approach[11,19]. All analyses were conducted using R version 4.2.3 (2023-03-15)[28]. The DTA estimation package is accessible via https://github.com/wbr-p/diagacc.

3. RESULTS

3.1 Study participants

The SIRS, depression, and epilepsy datasets included 36 (10 disease-free individuals), 33 (10 disease-free individuals), and 30 participants (6 disease-free individuals), respectively. For details on the participant's flow per diagnosis and demographic characteristics, see Appendix 6 and 7.

3.2 Test results per use case (intra-study evaluation)

We observed relevant differences within and across the different use cases for the three levels and time units (Figure 3 and Table 1).

278

279 **3.2.1 SIRS**

280 The DTA evaluation for the different time units estimated sensitivities of 84.8-93.6% on the time-
281 level, of 43.5-86.4% without modifying rules and 71.7-93.2% with modifying rules on the event-
282 level with blocks based on RS, of 62.0-86.3% on the event-level with blocks based on IT and RS,
283 and of 58.0-82.9% on the patient-time-level. Specificities of 94.4-97.6% on the time-level, of 68.1-
284 82.5% on the event-level without modifying rules and 90.5-97.6% with modifying rules on the
285 event-level with blocks based on RS, of 75.4-83.7% with blocks based on IT and RS, and of 71.6-
286 81.1% on the patient-time-level were estimated.

287

288 **3.2.2 Depression**

289 The DTA assessment for the different time units estimated sensitivities and specificities that were
290 ranging from 93.6% to 93.9% and from 97.5% to 98.3% on the time-level, respectively. On the
291 event-level with blocks based on RS, the unmodified sensitivity was 79.3% for all time units and
292 the unmodified specificities ranged between 29.0-98.4%, while the modified DTA estimated an
293 86.2% sensitivity for all time units and specificities of 69.4-98.4%. The DTA of the event-level
294 with blocks based on IT and RS estimated sensitivities of 80.6% for all time units and specificities
295 of 58.9-98.4%. On the patient-time-level, the sensitivity was 78.6% irrespective of the used time
296 unit, while the specificities ranged between 60.0% and 97.1%.

297

298 **3.2.3 Epilepsy**

299 The use case epilepsy estimated sensitivity and specificity of 80.5-98.7% and 98.1-98.5% on the
300 time-level, respectively. Sensitivities of 73.4-96.9% and specificities of 67.3-96.0% were estimated
301 on the event-level with blocks based on RS without modifying rules, while sensitivities of 71.7-

93.2% and specificities of 90.5-97.6% were estimated on the event-level with blocks based on RS after applying modifying rules. On the event-level with blocks based on IT and RS, DTA ranges of 74.2-97.1% sensitivities and 76.6-96.0% specificities were estimated, while sensitivities and specificities ranged between 55.8% and 96.0% and 58.8% and 95.0%, respectively, on the patient-time-level.

3.3 Test results across use cases (inter-study evaluation)

The evaluation across use cases showed that the highest DTAs, irrespective of used time unit and/or diagnosis, were estimated on the time-level, while the DTA on the event-level and patient-time-level were lower. The event-level with blocks based on RS showed that the unmodified DTA estimates were decreased compared to the DTA estimates after IT correction which sensitively depend on the chosen tolerance margin and/or %-correction rule. Moreover, the DTA estimates using ‘days’ as a time unit were closer to 100% than the DTA estimates using ‘hours’ or ‘minutes’ as time units. The number of observations decreased dramatically from the time-level to the event-level and/or patient-time-level which is somewhat mirrored by the DTA estimates.

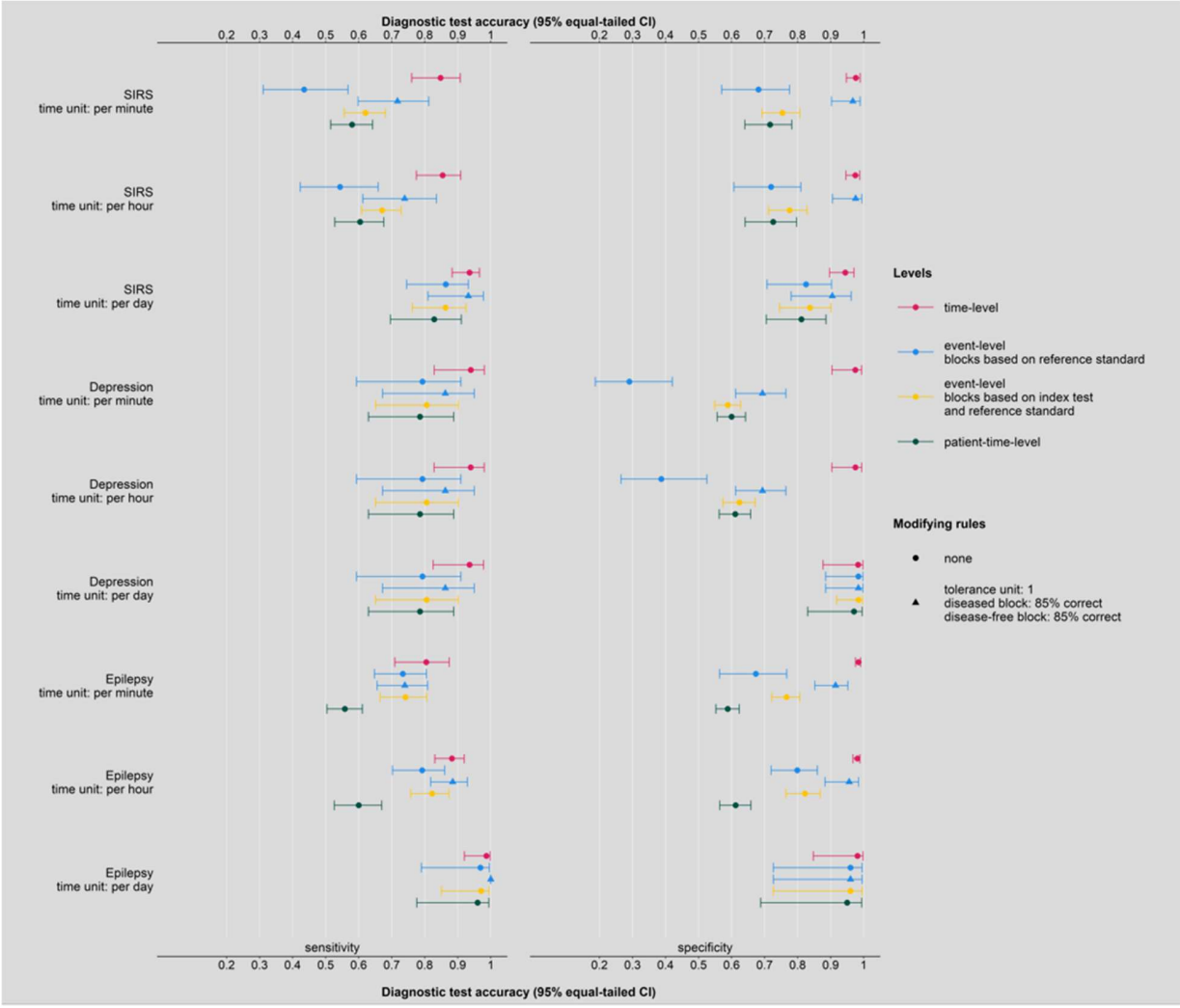


Figure 3: Summary of the diagnostic test accuracy of all three diagnoses stratified by the diagnostic test accuracy indices (i.e., sensitivity and specificity), by the estimation level (i.e., time-level, event-level, and patient-time-level), and by the time unit (i.e., minute, hour, and day).

Diagnosis	Time unit	Level	Sensitivity (95% CI)	Specificity (95% CI)	True Positive	False Positive	False Negative	True Negative
Systemic Inflammatory Response Syndrome (SIRS)	**Minute**	**Time**	84.8% (76.1-90.7%)	97.6% (94.8-98.9%)	84,243	12,315	15,124	500,847
		Event[1] *(blocks based on RS)*	43.5% (31.1-56.8%)	68.1% (57.0-77.5%)	20	29	26	62
		Event[2] *(blocks based on RS)*	71.7% (59.8-81.2%)	96.7% (90.2-98.9%)	33	3	19	88
		Event *(blocks based on IT and RS)*	62.0% (55.6-68.1%)	75.4% (69.2-80.6%)	49	33	30	101
		Patient-time	58.0% (51.5-64.2%)	71.6% (64.0-78.2%)	29	19	21	48
	Hour	**Time**	85.4% (77.5-90.9%)	97.5% (94.6-98.8%)	1,454	216	248	8,315
		Event[1] *(blocks based on RS)*	54.3% (42.3-65.9%)	72.0% (60.7-81.0%)	25	23	21	59
		Event[2] *(blocks based on RS)*	73.9% (61.2-83.6%)	97.6% (90.6-99.4%)	34	2	12	80
		Event *(blocks based on IT and RS)*	67.1% (60.9-72.8%)	77.6% (71.2-82.9%)	49	26	24	90
		Patient-time	60.4% (52.8-67.6%)	72.6% (64.1-79.7%)	29	17	19	45
	Day	**Time**	93.6% (88.3-96.6%)	94.4% (89.7-97.1%)	102	19	7	321
		Event[1] *(blocks based on RS)*	86.4% (74.5-93.2%)	82.5% (70.7-90.2%)	38	11	6	52
		Event[2] *(blocks based on RS)*	93.2% (81.0-97.8%)	90.5% (78.0-96.2%)	41	6	3	57
		Event *(blocks based on IT and RS)*	86.3% (76.3-92.5%)	83.7% (74.5-90.1%)	44	13	7	67
		Patient-time	82.9% (69.6-91.1%)	81.1% (70.5-88.6%)	29	10	6	43

Diagnosis	Time unit	Level	Sensitivity (95% CI)	Specificity (95% CI)	True Positive	False Positive	False Negative	True Negative
Depression	Minute	**Time**	93.9% (82.9-98.0%)	97.5% (90.4-99.4%)	728,122	75,338	46,926	2,886,594
		Event[1] *(blocks based on RS)*	79.3% (59.4-91.0%)	29.0% (18.7-42.0%)	23	44	6	18
		Event[2] *(blocks based on RS)*	86.2% (67.2-95.0%)	69.4% (61.2-76.4%)	25	19	4	43
		Event *(blocks based on IT and RS)*	80.6% (65.1-90.2%)	58.9% (54.9-62.7%)	29	44	7	63
		Patient-time	78.6% (63.0-88.8%)	60.0% (55.6-64.2%)	22	22	6	33
	Hour	**Time**	93.9% (82.8-98.0%)	97.5% (90.3-99.4%)	12,159	1,232	788	48,104
		Event[1] *(blocks based on RS)*	79.3% (59.4-91.0%)	38.7% (26.5-52.5%)	23	38	6	24
		Event[2] *(blocks based on RS)*	86.2% (67.2-95.0%)	69.4% (61.2-76.4%)	25	19	4	43
		Event *(blocks based on IT and RS)*	80.6% (65.1-90.2%)	62.4% (57.4-67.1%)	29	38	7	63
		Patient-time	78.6% (63.0-88.8%)	61.1% (56.2-65.8%)	22	21	6	33
	Day	**Time**	93.6% (82.5-97.8%)	98.3% (87.7-99.8%)	523	35	36	2,045
		Event[1] *(blocks based on RS)*	79.3% (59.4-91.0%)	98.4% (88.5-99.8%)	23	1	6	61
		Event[2] *(blocks based on RS)*	86.2% (67.2-95.0%)	98.4% (88.5-99.8%)	25	1	4	61
		Event *(blocks based on IT and RS)*	80.6% (65.1-90.2%)	98.4% (91.8-99.7%)	29	1	7	63
		Patient-time	78.6% (63.0-88.8%)	97.1% (83.0-99.6%)	22	1	6	33

Diagnosis	Time unit	Level	Sensitivity (95% CI)	Specificity (95% CI)	True Positive	False Positive	False Negative	True Negative
Epilepsy	Minute	**Time**	80.5% (70.9-87.4%)	98.5% (97.5-99.0%)	8,485	2,409	2,061	153,075
		Event[1] *(blocks based on RS)*	73.4% (64.8-80.5%)	67.3% (56.4-76.7%)	124	65	45	134
		Event[2] *(blocks based on RS)*	74.0% (65.6-80.9%)	91.5% (85.2-95.2%)	125	17	44	182
		Event *(blocks based on IT and RS)*	74.2% (66.5-80.6%)	76.6% (72.2-80.6%)	132	71	46	233
		Patient-time	55.8% (50.4-61.1%)	58.8% (55.2-62.3%)	24	21	19	30
	Hour	**Time**	88.2% (83.1-91.9%)	98.1% (96.8-98.9%)	277	47	37	2,422
		Event[1] *(blocks based on RS)*	79.2% (70.2-86.1%)	79.9% (72.0-86.1%)	103	32	27	127
		Event[2] *(blocks based on RS)*	88.5% (81.8-92.9%)	95.6% (88.3-98.4%)	115	7	15	152
		Event *(blocks based on IT and RS)*	82.2% (75.8-87.3%)	82.2% (76.4-86.8%)	125	35	27	162
		Patient-time	60.0% (52.6-67.0%)	61.2% (56.4-65.8%)	24	19	16	30
	Day	**Time**	98.7% (92.0-99.8%)	98.1% (84.8-99.8%)	74	1	1	52
		Event[1] *(blocks based on RS)*	96.9% (79.0-99.6%)	96.0% (72.6-99.5%)	31	1	1	24
		Event[2] *(blocks based on RS)*	100% (100-100%)	96.0% (72.6-99.5%)	32	1	0	24
		Event *(blocks based on IT and RS)*	97.1% (85.1-99.5%)	96.0% (72.6-99.5%)	33	1	1	24
		Patient-time	96.0% (77.6-99.4%)	95.0% (68.8-99.4%)	24	1	1	19

Diagnosis	Time unit	Level	Sensitivity (95% CI)	Specificity (95% CI)	True Positive	False Positive	False Negative	True Negative

[1] Event-level based on RS: No correction rules applied.
[2] Event-level based on RS: Application of ± 1 time point tolerance margin at start/end of RS episode and 85%-correction within diseased and disease-free blocks.

Table 1: Summary of the diagnostic test accuracy per diagnostic level (i.e., per time-level, per event-level, and per patient-time-level) per time unit (i.e., per minute, per hour, and per day) for the diagnoses 'Systemic Inflammatory Response Syndrome' (SIRS), depression, and epilepsy.

4. DISCUSSION

Our study shows that two features should be considered when presenting the DTA of an IT in the case of repeated test application and longitudinal data. These are the estimation level and the time unit which should always be chosen in accordance with diagnosis-specific characteristics and possible changes in the number of patient-related observations per cluster. An inappropriate feature in accordance with a specific research question and related estimand causes an increase in TP and TN observations. The selected diagnosis-specific choices of the estimation level and time unit show hereby a clear relation to the number of observations included in the DTA estimation. Because the time-level includes every single time point, the DTA estimation may be enriched with TP and TN observations. This is less problematic for the event-level that summarizes time points per diseased and disease-free block into a single, block-specific label; thus, fewer labels are included in the DTA estimation. This requires that the blocks are based on the diagnostic status of the RS so that the length of the blocks is fixed (i.e., not subject to the random variable IT). Moreover, if the time unit does not reflect well the diagnosis-specific characteristics, the DTA estimates may either have increased precision when using a small unit (i.e., increase in observations), or be distorted due to losing information as the unit was too large (e.g., epileptic seizures last only seconds to minutes which excludes using ‘days’ as time unit). In the last case, possible differences between the tests resulting in FN or FP labels may be lost. Estimand and statistical approach should be chosen appropriately so that they account for the longitudinal data format, because each of these features impacts the DTA estimation[18]. Using a simple approach not accounting for this specific data structure leads to a considerable overestimation of DTA when compared with what is relevant for clinical practice.

DTA can be reported using different levels. However, most studies reported their used analytical procedures and reporting level[29] rather intransparently; only few provided details on the estimation level and how it was constructed. For example, Wulff *et al.*[23] used the time-level (time unit: days) and the patient-time-level to present their CDSS's performance. Bode *et al.*[30] formed blocks that were labeled and the labels of the individual blocks were included in the DTA estimation (i.e., event-level with blocks based on IT and RS). Generally, various estimation levels can be used for the analysis and reporting of an IT's DTA, but researchers should carefully consider their research objective(s), related estimand(s), and potential differences of interpretation between the estimation levels, particularly in the context of longitudinal data. In the evaluation of longitudinal data, the time-level is always a good technical starting point, since the event-levels and the patient-time-level are based on the labeling of every time point so that they can be derived from the time-level DTA estimation. We recommend using the time-level when having a disease with short episodes (e.g., epileptic seizures), when the IT aims to predict a disease, or if the aim is to assess the IT's precision. The event-level with blocks based on RS can be used if the aim is to assess the IT's performance in a clinical setting (i.e., here the focus is on the periods that have correctly or incorrectly been classified by the IT) without having a constant decision to make (i.e., decision only required when IT changes its diagnostic status). The event-level with blocks based on IT and RS and patient-time-level are not suited for usage as discussed in the Methods. We recommend reporting multiple DTA estimations of a test using various level and time unit combinations while still considering related differences in interpretation.

The time unit, which was used for classifying a patient as either diseased or disease-free, also influences DTA estimation because it affects the number of labeled observations in the clusters. Many DTA studies do not provide sufficient information to identify if they used longitudinal data[8]

and/or their time unit. Studies that use longitudinal data should report their time unit and how they account for the inflation of the type I error in DTA estimation[31], which is caused by having repeated measures per patient. We identified few studies (e.g., [8,30,32–37]) that indicated or hinted at their used time unit. As with the estimation level, the used time unit also influences the interpretation and understanding of the DTA estimates[31]. In the previously presented examples, we show that the time unit has a critical implication on the IT's performance; hence, the IT's DTA estimation may be misleading if the time unit does not represent the disease-specific character (e.g., DTA of epilepsy reported with time unit 'days'). Using a short time unit has the advantage to increase precision, as the number of observations per cluster increases[38]. Additionally, we recommend being specific in the date and time classification of an episode to ensure an adequate evaluation. If, for example, both tests classify per day (i.e., starting at 00:00 am and finishing at 11:59 pm), then the effect on the DTA estimation, irrespective of time unit and estimation level, would be that the DTA estimates are identical. This is caused by equally inflating the number of observations included in the clusters in comparison to fewer numbers of observations.

Characteristics of the diagnosis must be considered even before performing the DTA estimation, as they determine the required time unit. The estimation level is somewhat unaffected, but researchers should select a level that best represents the research aim. As shown in the epilepsy example, the sensitivity and specificity estimates of all three estimation levels using 'days' as time unit differed barely. Other diseases which are characterized by medium to long episode periods and medium to long disease-free intervals between episodes, such as SIRS[39] or depression[40–42], could theoretically be assessed using any of the three time units. However, using 'minutes' as time unit significantly increases the number of observations; thus, the evaluation using an estimation tool could be slower due to the large number of observations, while also being more precise. We

suggest to only use a small time unit if the aim is to precisely and correctly assess the times of an episode start and end. The translation of observed DTA in a clinically meaningful DTA is often hampered when using small time units as it is inflated when compared to larger time units.

4.1 Limitations

All original datasets were collected with a defined study-specific purpose and modified to some extent (Appendices 4 and 5); thus, they are subject to a certain risk of data-generating pitfalls[43]. Especially, the depression and epilepsy datasets lacked information on IT and RS diagnoses; hence, IT and RS diagnoses were produced based on the available information in the datasets. Incorporation bias is most likely present in both datasets[44]. However, for this study's purpose it remains unconcerning because we aimed to demonstrate the problem in estimating an IT's DTA using longitudinal data. Note that the simulation of the depression data may likely not reflect a real-life situation (i.e., we expected a similar behavior of DTA estimates compared to the other diagnoses).

In this study, we assumed that the RSs perfectly diagnosed the diseases. Depending on the clinical setting, this might not be true, especially in situations where the DT is expected to alert clinicians before the RS becomes positive. Researchers should also keep in mind that the IT and/or RS can change over time (e.g., updated guidelines for diagnosis).

5. Conclusion

Using longitudinal data in a DTA study requires researchers to consider methodological choices and a clear pre-defined estimand early in the planning phase. Choices need to be made on the estimation level and the time unit considering diagnostic-specific characteristics as well as the related number of observations included in the DTA estimation. When reporting the DTA study's

findings, researchers should be transparent and state their rationale for the previously made choices. Researchers are not limited to reporting only one estimation level and/or time unit. As a next step, these methodological approaches could be improved by using a nonparametric approach that incorporates the structured correlation of the time series evaluation as well as other characteristics of a real-life dataset (e.g., missing values).

LIST OF ABBREVIATIONS

Abbreviation	Definition
CDSS	Clinical decision support system
CI	Confidence interval
DT	Diagnostic test
DTA	Diagnostic test accuracy
EEG	Electroencephalogram
FN	False negative
FP	False positive
IT	Index test
RS	Reference standard
TN	True negative
TP	True positive

DECLARATIONS

Ethics approval and consent to participate

For this study, we used only publicly available anonymized datasets. No approval by an institutional review board or regional review board was required as this study has no direct implications for their health and wellbeing.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the German Federal Ministry of Health via the ELISE project (grant number 2520DAT66C) and by the German Research Foundation (DFG; grant number 499188607).

Registration and accessibility of the study protocol

This work was neither registered nor did we publish a study protocol because it is a study demonstrating theoretical approaches with use cases from clinical practice and not a diagnostic test accuracy study in itself.

Availability of data and materials

The original datasets can be accessed via the data owners (see “2.3 The datasets”); the modified-labeled datasets including the R-Code for the dataset modifications can be accessed via

https://zivgitlab.uni-muenster.de/ruebsame/dta_longitudinal_data_methods. All rights of the modified-labeled datasets remain with the data owners of this publication. The R package "diagacc" can be accessed via https://github.com/wbr-p/diagacc.

Acknowledgements

We thank Maria Stark (Department of Medical Biometry and Epidemiology, University Medical Center Hamburg, Hamburg, Germany), Jürgen Wellmann (Institute of Epidemiology and Social Medicine, University of Münster, Münster, Germany), and Johannes B. Reitsma (Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands) for their valuable input in understanding and dealing with the challenge of diagnostic test accuracy estimation when using longitudinal data.

Authors' contributions (CRediT Taxonomy)

Julia Böhnke: Conceptualization (equal); data curation (equal); formal analysis (lead); investigation (equal); methodology (lead); project administration (equal); resources (equal); visualization (lead); writing – original draft preparation (lead); writing - review and editing (lead). **Antonia Zapf:** Conceptualization (equal); data curation (equal); formal analysis (equal); investigation (equal); methodology (equal); project administration (equal); resources (equal); supervision (equal); visualization (equal); writing – original draft preparation (equal); writing – review and editing (equal). **Katharina Kramer:** Conceptualization (equal); data curation (equal); formal analysis (equal); investigation (equal); methodology (equal); project administration (equal); resources (equal); supervision (equal); visualization (equal); writing – original draft preparation (equal); writing – review and editing (equal). **Philipp Weber:** Writing – analysis program (lead); Writing – review and editing (equal). **ELISE Study Group:** Writing – review and editing (equal).

475 **André Karch:** Conceptualization (equal); data curation (equal); formal analysis (equal);
476 investigation (equal); methodology (equal); project administration (equal); resources (equal);
477 supervision (equal); visualization (equal); writing – original draft preparation (equal); writing –
478 review and editing (equal). **Nicole Rübsamen:** Conceptualization (equal); data curation (equal);
479 formal analysis (equal); investigation (equal); methodology (equal); project administration (equal);
480 resources (equal); supervision (equal); visualization (equal); writing – original draft preparation
481 (equal); writing – review and editing (equal).

REFERENCES

1. Definition of diagnostic test - NCI Dictionary of Cancer Terms - National Cancer Institute. Accessed March 31, 2022. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/diagnostic-test

2. Leeflang MMG, Allerberger F. How to: evaluate a diagnostic test. *Clin Microbiol Infect*. 2019;25(1):54-59. doi:10.1016/J.CMI.2018.06.011

3. European Parliament and Council of the European Union. *Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on Medical Devices, Amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and Repealing Council Directives 90/385/EEC and 93/42/EE.*; 2017:1-175. Accessed April 5, 2022. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32017R0745

4. Hoyer A, Zapf A. Studies for the evaluation of diagnostic tests:part 28 of a series on evaluation of scientific publications. *Dtsch Arztebl Int*. 2021;118(33-34):550-560. doi:10.3238/ARZTEBL.M2021.0224

5. Chassé M, Fergusson DA. Diagnostic Accuracy Studies. *Semin Nucl Med*. 2019;49(2):87-93. doi:10.1053/J.SEMNUCLMED.2018.11.005

6. Sitch AJ, Dekkers OM, Scholefield BR, Takwoingi Y. Introduction to diagnostic test accuracy studies. *Eur J Endocrinol*. 2021;184(2):E5-E9. doi:10.1530/EJE-20-1239

7. Miller DC, Dunn RL, Wei JT. Assessing the Performance and Validity of Diagnostic Tests and Screening Programs. *Clin Res Methods Surg*. Published online 2006:157-174. doi:10.1007/978-1-59745-230-4_10

8. Böhnke J, Varghese J, Karch A, et al. Systematic review identifies deficiencies in reporting of diagnostic test accuracy among clinical decision support systems. *J Clin*

Epidemiol. 2022;151:171-184. doi:10.1016/J.JCLINEPI.2022.08.003

9. Ochodo EA, De Haan MC, Reitsma JB, Hooft L, Bossuyt PM, Leeflang MMG. Overinterpretation and misreporting of diagnostic accuracy studies: evidence of "spin." *Radiology*. 2013;267(2):581-588. doi:10.1148/RADIOL.12120527

10. Genders TSS, Spronk S, Stijnen T, Steyerberg EW, Lesaffre E, Hunink MGM. Methods for calculating sensitivity and specificity of clustered data: A tutorial. *Radiology*. 2012;265(3):910-916. doi:10.1148/RADIOL.12120509/ASSET/IMAGES/LARGE/120509T02.JPEG

11. Lange K. *Nichtparametrische Analyse Diagnostischer Gütemaße Bei Clusterdaten* [Dissertation]. Georg-August-University Göttingen, Germany; 2011. doi:10.53846/GOEDISS-3538

12. Gönen M, Panageas KS, Larson SM. Statistical Issues in Analysis of Diagnostic Imaging Experiments with Multiple Observations per Patient. *Radiology*. 2001;221(3):763-767. doi:10.1148/RADIOL.2212010280

13. Mondol MH, Rahman MS. Bias-reduced and separation-proof GEE with small or sparse longitudinal binary data. *Stat Med*. 2019;38(14):2544-2560. doi:10.1002/SIM.8126

14. Miao Z, Tang LL, Yuan A. Comparative study of statistical methods for clustered ROC data: nonparametric methods and multiple outputation methods. *Biostat Epidemiol*. 2021;5(2):169-188. doi:10.1080/24709360.2021.1880224

15. European Medicines Agency. *ICH E9 (R1) Addendum on Estimands and Sensitivity Analysis in Clinical Trials to the Guideline on Statistical Principles for Clinical Trials*.; 2020. Accessed January 24, 2023. https://www.ema.europa.eu/en/documents/scientific-guideline/draft-ich-e9-r1-addendum-estimands-sensitivity-analysis-clinical-trials-guideline-statistical_en.pdf

16. Lawrance R, Degtyarev E, Griffiths P, et al. What is an estimand & how does it relate to quantifying the effect of treatment on patient-reported quality of life outcomes in clinical trials? *J patient-reported outcomes*. 2020;4(1). doi:10.1186/S41687-020-00218-5

17. Pohl M, Baumann L, Behnisch R, Kirchner M, Krisam J, Sander A. Estimands - A Basic Element for Clinical Trials: Part 29 of a Series on Evaluation of Scientific Publications. *Dtsch Arztebl Int*. 2021;118(51-52):883. doi:10.3238/ARZTEBL.M2021.0373

18. Fierenz A, Rackow B, Zapf A. GMS | 67. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS), 13. Jahreskongress der Technologie- und Methodenplattform für die vernetzte medizinische Forschung e. V. (TMF). In: *The Estimand Framework in Diagnostic Studies*. German Medical Science GMS Publishing House; 2022. doi:10.3205/22GMDS085

19. Konietschke F, Brunner E. Nonparametric analysis of clustered data in diagnostic trials: Estimation problems in small sample sizes. *Comput Stat Data Anal*. 2009;53(3):730-741. doi:10.1016/J.CSDA.2008.08.006

20. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ*. 2015;351. doi:10.1136/BMJ.H5527

21. Cui Y, Konietschke F, Harrar SW. The nonparametric Behrens-Fisher problem in partially complete clustered data. *Biom J*. 2021;63(1):148-167. doi:10.1002/BIMJ.201900310

22. Lange K, Brunner E. Sensitivity, specificity and ROC-curves in multiple reader diagnostic trials—A unified, nonparametric approach. *Stat Methodol*. 2012;9(4):490-500. doi:10.1016/J.STAMET.2011.12.002

23. Wulff A, Montag S, Rübsamen N, et al. Clinical evaluation of an interoperable clinical decision-support system for the detection of systemic inflammatory response syndrome in critically ill children. *BMC Med Inform Decis Mak*. 2021;21(1). doi:10.1186/S12911-021-

01428-7

24. Wulff A, Mast M, Bode L, Rathert H, Jack T. Towards an Evolutionary Open Pediatric Intensive Care Dataset in the ELISE Project. *Stud Health Technol Inform*. 2022;295. doi:10.3233/SHTI220670

25. Wulff A, Mast M, Bode L, et al. ELISE - An open pediatric intensive care data set. Published 2022. Accessed August 25, 2022. https://leopard.tu-braunschweig.de/receive/dbbs_mods_00070468

26. Garcia-Ceja E, Riegler M, Jakobsen P, et al. Depresjon: A motor activity database of depression episodes in unipolar and bipolar patients. *Proc 9th ACM Multimed Syst Conf MMSys 2018*. Published online June 12, 2018:472-477. doi:10.1145/3204949.3208125

27. Shoeb A. *Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment*. Massachusetts Institute of Technology; 2009. Accessed July 19, 2022. https://dspace.mit.edu/handle/1721.1/54669

28. R Core Team. R: A Language and Environment for Statistical Computing. Published online 2021. https://www.r-project.org/

29. Westwood M, Joore M, Grutters J, et al. Contrast-enhanced ultrasound using SonoVue® (sulphur hexafluoride microbubbles) compared with contrast-enhanced computed tomography and contrast-enhanced magnetic resonance imaging for the characterisation of focal liver lesions and detection of liver met. *Health Technol Assess (Rockv)*. 2013;17(16):7-243. doi:10.3310/HTA17160

30. Bode L, Schamer S, Böhnke J, et al. Tracing the Progression of Sepsis in Critically Ill Children: Clinical Decision Support for Detection of Hematologic Dysfunction. *Appl Clin Inform*. 2022;13(5). doi:10.1055/A-1950-9637

31. Parsons NR, Teare MD, Sitch AJ. Unit of analysis issues in laboratory-based research.

Elife. 2018;7. doi:10.7554/ELIFE.32486

32. Dewan M, Muthu N, Shelov E, et al. Performance of a Clinical Decision Support Tool to Identify PICU Patients at High Risk for Clinical Deterioration. *Pediatr Crit Care Med*. 2020;21(2):129-135. doi:10.1097/PCC.0000000000002106

33. Nagori A, Dhingra LS, Bhatnagar A, Lodha R, Sethi T. Predicting Hemodynamic Shock from Thermal Images using Machine Learning. 2019;9(1):1-9. Accessed May 9, 2022. https://pubmed.ncbi.nlm.nih.gov/30643187/

34. Calvert JS, Price DA, Chettipally UK, et al. A computational approach to early sepsis detection. *Comput Biol Med*. 2016;74:69-73. doi:10.1016/J.COMPBIOMED.2016.05.003

35. Wulff A, Haarbrandt B, Tute E, Marschollek M, Beerbaum P, Jack T. An interoperable clinical decision-support system for early detection of SIRS in pediatric intensive care using openEHR. *Artif Intell Med*. 2018;89:10-23. doi:10.1016/J.ARTMED.2018.04.012

36. Wulff A, Montag S, Steiner B, et al. CADDIE2-evaluation of a clinical decision-support system for early detection of systemic inflammatory response syndrome in paediatric intensive care: study protocol for a diagnostic study. *BMJ Open*. 2019;9(6). doi:10.1136/BMJOPEN-2019-028953

37. Wulff A, Montag S, Rübsamen N, et al. Clinical evaluation of an interoperable clinical decision-support system for the detection of systemic inflammatory response syndrome in critically ill children. *BMC Med Inform Decis Mak*. 2021;38(1):219-226. doi:10.1186/s12911-021-01428-7

38. Hess AS, Shardell M, Johnson JK, et al. Methods and recommendations for evaluating and reporting a new diagnostic test. *Eur J Clin Microbiol Infect Dis*. 2012;31(9). doi:10.1007/S10096-012-1602-1

39. Chakraborty RK, Burns B. Systemic Inflammatory Response Syndrome. StatPearls.

602 Published 2022. Accessed August 17, 2022.
603 https://www.ncbi.nlm.nih.gov/books/NBK547669/
604 40. Chand SP, Arif H. *Depression*. StatPearls Publishing; 2022. Accessed November 15, 2022.
605 https://www.ncbi.nlm.nih.gov/books/NBK430847/
606 41. Goodwin G. Depression. In: Castle D, Coghill D, eds. *Comprehensive Men's Mental*
607 *Health*. Cambridge University Press; 2021:128-138. doi:10.1017/9781108646765.013
608 42. Strunk DR, Pfeifer BJ, Ezawa ID. Depression. In: Wenzel A, ed. *Handbook of Cognitive*
609 *Behavioural Therapy: Applications., Vol. 2.* American Psychological Association; 2021:3-
610 31. doi:http://dx.doi.org/10.1037/0000219-001
611 43. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical
612 methods. *Stat Med*. 2019;38(11):2074-2102. doi:10.1002/SIM.8086
613 44. Catalogue of Bias Collaboration. Incorporation bias. In: Plüddemann A, McCall M, eds.
614 *Scakett Catalogue of Biases 2019*. ; 2019. Accessed November 21, 2022.
615 https://catalogofbias.org/biases/incorporation-bias/
616

ONLINE SUPPLEMENTARY MATERIALS