Journal of Clinical Epidemiology 169 (2024) 111314

Journal of Clinical Epidemiology

ORIGINAL RESEARCH

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

Julia Böhnke[a,*], Antonia Zapf[c], Philipp Weber[b], ELISE Study Group[1], André Karch[a,2], Nicole Rübsamen[a,2]

[a]Institute of Epidemiology and Social Medicine, University of Münster, Albert-Schweitzer-Campus 1, 48149 Münster, Germany
[b]Department of Medical Biometry and Epidemiology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
[c]Mathematical Statistics and Artificial Intelligence in Medicine, University of Augsburg, Augsburg, Germany

Accepted 26 February 2024; Published online 1 March 2024

## Abstract

**Objectives:** In this study, we evaluate how to estimate diagnostic test accuracy (DTA) correctly in the presence of longitudinal patient data (ie, repeated test applications per patient).

**Study Design and Setting:** We used a nonparametric approach to estimate the sensitivity and specificity of three tests for different target conditions with varying characteristics (ie, episode length and disease-free intervals between episodes): 1) systemic inflammatory response syndrome ($n = 36$), 2) depression ($n = 33$), and 3) epilepsy ($n = 30$). DTA was estimated on the levels '*time*', '*block*', and '*patient-time*' for each diagnosis, representing different research questions. The estimation was conducted for the time units per minute, per hour, and per day.

**Results:** A comparison of DTA per and across use cases showed variations in the estimates, which resulted from the used level, the time unit, the resulting number of observations per patient, and the diagnosis-specific characteristics. Intra- and inter-use-case comparisons showed that the time-level had the highest DTA, particularly the larger the time unit, and that the patient-time-level approximated 50% sensitivity and specificity.

**Conclusion:** Researchers need to predefine their choices (ie, estimation levels and time units) based on their individual research aims, estimands, and diagnosis-specific characteristics of the target outcomes to make sure that unbiased and clinically relevant measures are communicated. In cases of uncertainty, researchers could report the DTA of the test using more than one estimation level and/or time unit.

*Keywords:* Diagnostic study; Diagnostic test accuracy; Longitudinal study; Data cluster; Nonparametric method; Estimation level

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].

[1] ELISE study group members: Louisa Bode (Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Hannover, Germany), Marcel Mast (Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Hannover, Germany), Antje Wulff (Big Data in Medicine, Department of Health Services Research, School of Medicine and Health Sciences, Carl von Ossietzky University Oldenburg, Oldenburg, Germany; Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Hannover, Germany), Michael Marschollek (Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Hannover, Germany), Sven Schamer (Department of Pediatric Cardiology and Intensive Care Medicine, Hannover Medical School, Hannover Germany), Henning Rathert (Department of Pediatric Cardiology and Intensive Care Medicine, Hannover Medical School, Hannover Germany), Thomas Jack (Department of Pediatric Cardiology and Intensive Care Medicine, Hannover Medical School, Hannover Germany), Philipp Beerbaum (Department of Pediatric Cardiology and Intensive Care Medicine, Hannover Medical School, Hannover Germany), Nicole Rübsamen (Institute of Epidemiology and Social Medicine, University of Münster, Münster, Germany), Julia Böhnke (Institute of Epidemiology and Social Medicine, University of Münster, Münster, Germany), André Karch (Institute of Epidemiology and Social Medicine, University of Münster, Münster, Germany), Pronaya Prosun Das (Research Group Bioinformatics, Fraunhofer Institute for Toxicology and Experimental Medicine, Hannover, Germany), Lena Wiese (Research Group Bioinformatics, Fraunhofer Institute for Toxicology and Experimental Medicine, Hannover, Germany), Christian Groszweski-Anders (medisite GmbH, Hannover, Germany), Andreas Haller (medisite GmbH, Hannover, Germany), Torsten Frank (medisite GmbH, Hannover, Germany)

[2] These authors contributed equally to this work.

* Corresponding author. Institute of Epidemiology and Social Medicine University of Münster, Albert-Schweitzer-Campus 1, 48149 Münster, Germany.

*E-mail address:* boehnkej@uni-muenster.de (J. Böhnke).

https://doi.org/10.1016/j.jclinepi.2024.111314
0895-4356/

**What is new?**

**Key findings**

- Diagnostic test accuracy (DTA) estimation in longitudinal study settings (ie, repeated test applications per subject) requires appropriate methodological approaches that account for the clustering of the data.

**What this adds to what was known?**

- Use cases showed that DTA estimates differed depending on the estimation level (ie, time-level, block-level, and/or patient-time-level) and time unit (eg, per minute, per hour, per day, etc.); they were even misleading if they did not reflect the diagnosis-specific characteristics (eg, epilepsy seizures last seconds to a few minutes, therefore the time unit 'day' is inadequate as small differences between the index test and the reference standard may not be captured).

**What is the implication and what should change now?**

- Researchers should preselect their choices of estimation level and time unit in accordance with diagnosis-specific characteristics of the target condition for the DTA estimation. If possible, multiple reporting options are desirable.

## 1. Introduction

A diagnostic test can be any device (eg, biomarker quantification, magnetic resonance imaging, etc.) [1−3] with which healthcare professionals classify a condition (eg, diseased vs disease-free) [1−6] and make an informed decision based on the test's result. Each test is required to be assessed for its diagnostic test accuracy (DTA) before its usage in practice [7]. Any diagnostic test should provide a correct classification of the presence or absence of a condition (ie, *true positive* [TP], *true negative* [TN]) while being safe and effective [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] and/or a country's health care system [2].

The diagnostic validity of the diagnostic test (referred to as the *index test*) is best assessed in a DTA study using an established *reference standard* as the ground truth [5,7]. To minimize potential influences, test read-outs should be blinded to each other and performed without time delay [2,5]. Information on test performance is usually reported in terms of sensitivity and specificity (Table 1).

Lately, researchers have shown that many DTA studies are of low quality, do not necessarily represent the situation of interest, and/or are associated with a considerable risk of bias [8,9]. Consequently, the diagnostic test under review might not be used in practice or the research may be involuntarily distorted [9,10]. Particularly, repeated measurements per patient require adequate DTA analysis approaches as the within-person correlation can inflate the diagnostic test's uncorrected accuracy compared to only including a single measurement per patient [11−13]. A systematic review highlighted that most DTA studies did not report sufficient information on the usage of or adjustment for longitudinal data (ie, repeated measurements per patients with disease-free and/or diseased intervals) in the DTA estimation [8]. Those that accounted for longitudinal data used various methods to adjust their DTA estimates [14].

When evaluating repeated diagnostic tests on the same person, treatment effects must also be considered. An early intervention may hinder the condition's onset, while treatment after diagnosis may cause a health improvement. An a priori definition of the estimand that is the target for a DTA estimation to address the scientific question of interest posed by the study objective [15−18] is, therefore, necessary.

This study 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 DTA estimates at three estimation levels (ie, time-level, block-level, and patient-time-level), and
- introducing a nonparametric estimation method [11,19].

## 2. Methods

We report this study in accordance with the Standard for Reporting Diagnostic Accuracy guideline [20] (Appendix 1). We use the following nomenclature: A "time unit" is chosen by the researcher, that is, diagnosis assessment every minute/hour/day. A "time point" refers to a specific minute/hour/day within the longitudinal setting. 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 (Figs 1−3) determining an index test's performance using longitudinal data. More elaborate descriptions, including step-by-step labeling instructions, are presented in Appendix 2.

### 2.1.1. Time-level

The time-level provides a label for every time point. This level's estimand is the diagnostic status (ie, target

**Table 1.** Key terminology of diagnostic test accuracy and diagnostic test accuracy studies

| Terminology | Description |
|---|---|
| Index test | The test of interest is called the index test. It can be "any medical device that is a reagent, reagent product, calibrator, control material, kit, instrument, apparatus, piece of equipment, software or system, whether used alone or in combination [...] for the purpose of providing information [...] concerning a physiological or pathological process or state." [3] |
| Reference standard | "This is the test used to define the target condition, and the underlying assumption is that it reflects the truth. By design, the reference standard is assumed to be flawless. The reference standard sets the reference, and sensitivity and specificity are expressed as the proportion of reference standard positives with a positive index test result, and the proportion of reference standard negatives with a negative index test result, respectively. It is therefore impossible to show that an index test is better than the reference standard, even if this would be the case in reality." [2] |
| True positive (TP) | At a given time point, both the index test and the reference standard detect the occurrence of the target condition. |
| True negative (TN) | At a given time point, both the index test and the reference standard do not detect the occurrence of the target condition. |
| False positive (FP) | At a given time point, the index test detects the occurrence of the target condition, but the reference standard does not. |
| False negative (FN) | At a given time point, the reference standard detects the occurrence of the target condition, but the index test does not. |
| Sensitivity | The probability of a positive index test given that the reference standard detects the occurrence of the target condition, estimated by $\sum TP/(\sum TP + \sum FN)$. |
| Specificity | The probability of a negative index test given that the reference standard does not detect the occurrence of the target condition, estimated by $\sum TN/(\sum TN + \sum FP)$. |
| Positive Predictive Value (PPV) | The probability of having the target condition given a positive index test result, estimated by $\sum TP/(\sum TP + \sum FP)$. |
| Negative Predicative Value (NPV) | The probability of not having the target condition given a negative index test result, estimated by $\sum TN/(\sum FN + \sum TN)$. |
| Diagnostic accuracy | The proportion of all test results, both positive and negative, that is correctly identified by the index test given the reference standard diagnostic health status (i.e., the true diagnostic health status), estimated by $(\sum TP + \sum TN)/(\sum TP + \sum FN + \sum FP + \sum TN)$. |
| Receiver Operating Characteristic (ROC) Curve | Expresses the relationship between the sensitivity and the specificity by plotting the true-positive rate (sensitivity) against the FP rate (1 - specificity) over a range of cut-off values. The ROC curve of a test that discriminates well is crowed toward the upper-left corner of the plane. |
| Area under the Curve (AUC) | Provides an aggregate measure of performance across all possible cut-off values. One way of interpreting AUC is as the probability that the index test for a random individual with the target condition is higher than for a random individual without the target condition. |

condition present/absent) per time point without any aggregation.

*2.1.2. Block-level*

The block-level aggregates consecutive, labeled time points based on diagnostic status. This level requires that the estimand is a change in the diagnostic status.

*2.1.2.1. Blocks based on reference standard.* The time point at which the reference standard 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 reference standard is assumed to be known, while the result of the index test is a random variable that follows a Bernoulli distribution.

For DTA estimation, the time point labels per block are summarized into one single label which is included in the DTA estimation. FP and FN labels overrule TP and TN labels, ie, this labeling penalizes any differences between the diagnostic tests. We can control for this by applying modifying rules, eg, applying a clinician-based tolerance margin rule, so that if the index test starts or ends within the tolerance margin of the reference standard, the index test's diagnostic status at the specific time points is changed in accordance with the reference standard's diagnostic status (ie, no "punishment" if the index test starts and/or ends too early or too late). However, if the index test starts or ends outside of the tolerance margin, the index test's diagnostic status of these specific time points remains unchanged. A %-correctness rule can also be applied according to which the index test's diagnostic status per patient is corrected in accordance with the reference standard's diagnostic status if at minimum $P_{diseased}\%$ of single time points per a diseased block and at minimum

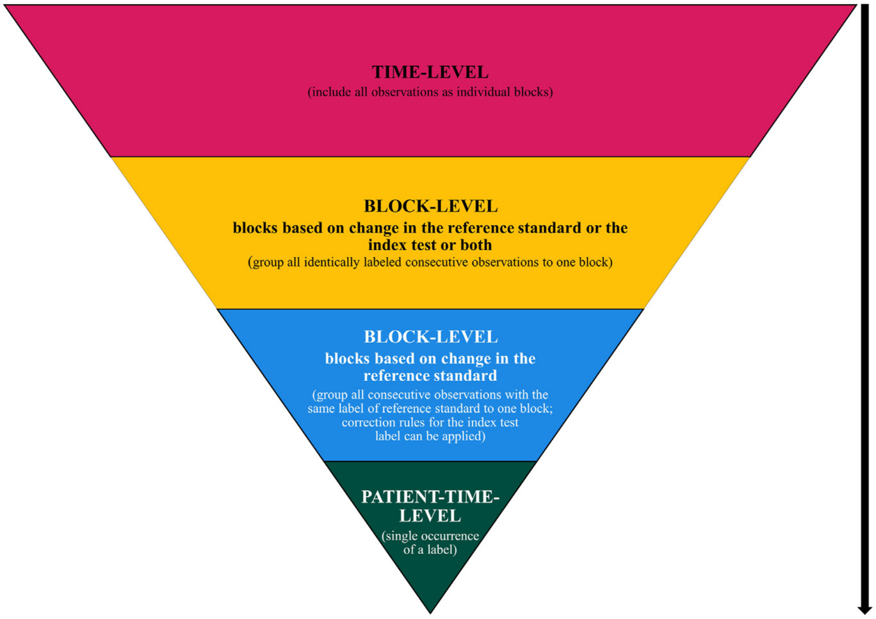

TIME-LEVEL
(include all observations as individual blocks)
BLOCK-LEVEL
blocks based on change in the reference standard or the index test or both
(group all identically labeled consecutive observations to one block)
BLOCK-LEVEL
blocks based on change in the reference standard
(group all consecutive observations with the same label of reference standard to one block; correction rules for the index test label can be applied)
PATIENT-TIME-LEVEL
(single occurrence of a label)
Number of subunits

**Figure 1.** Visualization of the data structure and its subunits that are included in the diagnostic test accuracy estimation. Two options for the block-level are presented that differ regarding their groupings of labeled time points into blocks. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

$P_{disease-free}\%$ of single time points per a disease-free block are correctly classified. The $P$'s are diagnosis-specific. For our analysis, we used a tolerance margin of ±1 time interval around the reference standard's disease episode start and end and an 85% correction rule for diseased and disease-free blocks (Appendix 3).

*2.1.2.2. Blocks based on index test and reference standard.* Each new block starts and ends with a change in the diagnostic status of the index test and/or the reference standard and is given a single summary label that is used for the DTA estimation. Modifying rules can be applied. With this definition, the results of both diagnostic tests are random variables, which violate 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 or at maximum four labels to the DTA estimation. This level's estimand is the occurrence of the 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 is a biased estimate of 50% sensitivity and 50% specificity.

## 2.2. Nonparametric approach for DTA estimation

The DTA can be estimated using a nonparametric approach [11,19] which is robust and reliable when accounting for intra- and intercluster correlations [21] without having to assume specific dependence structures and distributions within a cluster. It categorizes the patients into three clusters (Appendix 4), regardless of the individual participant's number of repeated measurements [19,21]:

- 'Absent' ($ic_0$): Patient was consistently disease-free during the total observation period.
- 'Present' ($ic_1$): Patient was consistently diseased during the total observation period.
- 'Mix' ($c$): Patient experienced diseased and disease-free phases during the total observation period.

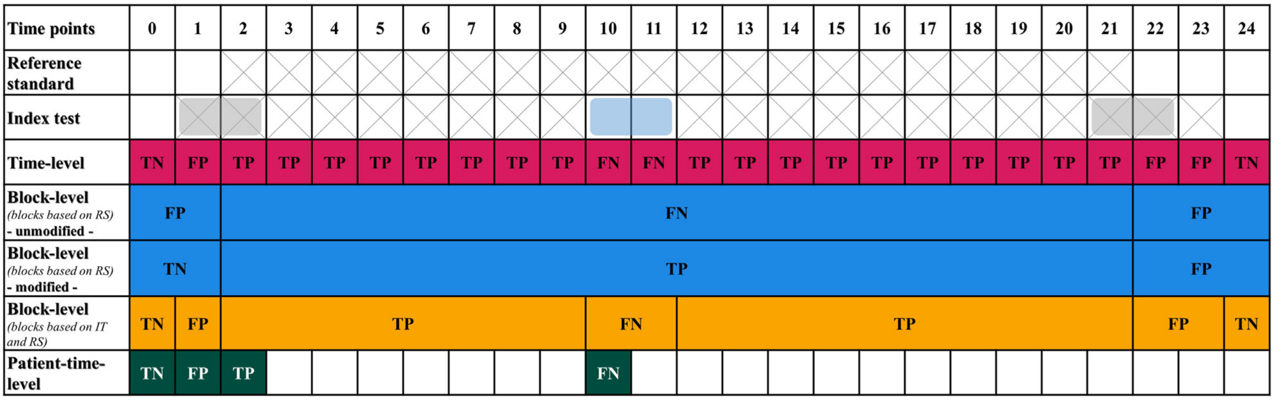
This method uses a unified nonparametric model to estimate the sensitivity and specificity accounting for the clustered longitudinal data [11,22]. It applies a nonparametric rank statistic using the weighted estimation strategy (ie, weighting by the size of the cluster) [11,21]. This allows assigning 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, we refer to [11,19].

## 2.3. Use cases

We used three publicly available datasets as use cases (Table 2) to show the application of our proposed methods. The dataset descriptions, labeling, and information on the diagnostic tests are presented in Appendices 2—3 and 5. The use cases were representative examples for distinctive target conditions that varied in episode and disease-free interval lengths and their respective frequencies (Table 2).

### 2.3.1. SIRS dataset

The SIRS dataset includes 168 pediatric patients. All participants were consecutively recruited at a single study



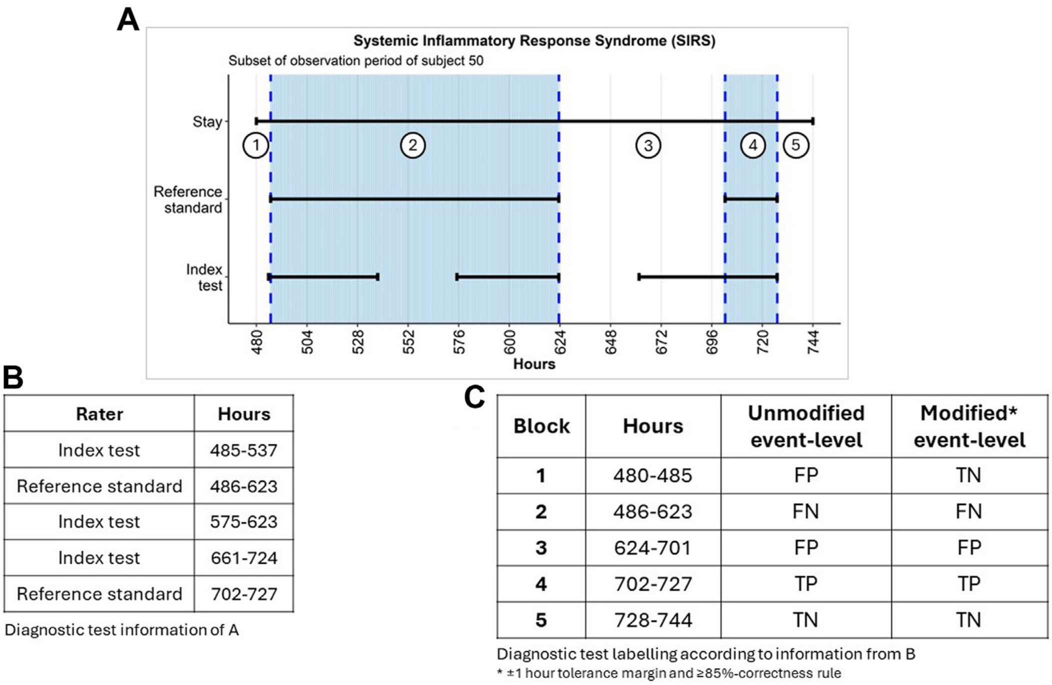
| Time points | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Reference standard | | | ☒ | ☒ | ☒ | ☒ | ☒ | ☒ | ☒ | ☒ | ☒ | ☒ | ☒ | ☒ | ☒ | ☒ | ☒ | ☒ | ☒ | ☒ | ☒ | ☒ | | | |
| Index test | | ☒ | ☒ | ☒ | ☒ | ☒ | ☒ | ☒ | ☒ | ☒ | ▭ | ▭ | ☒ | ☒ | ☒ | ☒ | ☒ | ☒ | ☒ | ☒ | ☒ | ☒ | ☒ | ☒ | |
| Time-level | TN | FP | TP | TP | TP | TP | TP | TP | TP | TP | FN | FN | TP | TP | TP | TP | TP | TP | TP | TP | TP | TP | FP | FP | TN |
| Block-level (blocks based on RS) - unmodified - | FP | | FN | | | | | | | | | | | | | | | | | | | | FP | | |
| Block-level (blocks based on RS) - modified - | TN | | TP | | | | | | | | | | | | | | | | | | | | FP | | |
| Block-level (blocks based on IT and RS) | TN | FP | TP | | | | | | | | FN | | TP | | | | | | | | | | FP | | TN |
| Patient-time-level | TN | FP | TP | | | | | | | | FN | | | | | | | | | | | | | | |

RS = reference standard, IT = index test, TP = true positive, FP = false positive, FN = false negative, TN = true negative

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.
- *Block-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.
- *Block-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), three false positive (FP), two false negative (FN), and two true negative (TN) observations to the DTA estimation. The DTA estimation of the block-level using blocks based on the reference standard (modified) adds 1 TP, 1 FP, and 1 TN to the DTA estimation, while the block-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. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)



| Rater | Hours |
|---|---|
| Index test | 485-537 |
| Reference standard | 486-623 |
| Index test | 575-623 |
| Index test | 661-724 |
| Reference standard | 702-727 |

Diagnostic test information of A

| Block | Hours | Unmodified event-level | Modified* event-level |
|---|---|---|---|
| 1 | 480-485 | FP | TN |
| 2 | 486-623 | FN | FN |
| 3 | 624-701 | FP | FP |
| 4 | 702-727 | TP | TP |
| 5 | 728-744 | TN | TN |

Diagnostic test labelling according to information from B
* ±1 hour tolerance margin and ≥85%-correctness rule

**Figure 3.** Subset (hours 488—744) of the full study period of a patient in the SIRS dataset. The black lines indicate the presence of the target condition at the specific time points. A total of five blocks according to the reference standard were summarized (ie, block 1: hours 480—485; block 2: hours 486—623; block 3: hours 624—701; block 4: hours 702—727; block 5: hours 728—744). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

**Table 2.** Demographic characteristics of participants (ie, information on the eligible and included participants) and summary of the test settings per modified diagnostic dataset

| Demographic characteristics | SIRS[a] dataset($n = 36$) | Depression dataset ($n = 33$) | Epilepsy dataset ($n = 30$) |
|---|---|---|---|
| Main characteristic of target condition | Medium-to-long episode and disease-free interval length (measured in hours), with a low frequency of recurrence | Medium-to-long episode and disease-free interval length (measured in days), with a medium frequency of recurrence | Short-to-medium episode and disease-free interval length (measured in minutes), with a high frequency of recurrence |
| Cohort | Pediatric intensive care patients, Hannover Medical School, Germany | Psychiatric adults, Haukeland University Hospital, Norway | Pediatric and young adults, Boston Children's Hospital, USA |
| Age | | | |
| Years at enrollment | 0–17 | 20–69 | 1–22 |
| Missing | 0 | 0 | 6 |
| Sex | | | |
| Male | 22 (61.1%) | 17 (51.5%) | 5 (16.7%) |
| Female | 14 (38.9%) | 16 (48.5%) | 17 (56.7%) |
| Missing | 0 (0.0%) | 0 (0.0%) | 8 (26.7%)[b] |
| Disease status | | | |
| Diseased (at least one episode) | 26 (72.2%) | 23 (69.7%) | 24 (80.0%) |
| Disease-free at any time | 10 (27.8%) | 10 (30.3%) | 6 (20.0%) |
| Length of observation period | | | |
| Total | 10,233 h | 62,932 h | 2783 h |
| Range per patient (Min–Max) | 21–1122 h | 1583–2399 h | 50–233 h |
| Number of episodes per patient | | | |
| 0 (ie, disease-free) | 10 (27.8%) | 10 (30.3%) | 6 (20.0%) |
| 1 | 13 (36.1%) | 17 (51.5%) | 0 (0.0%) |
| 2 | 10 (27.8%) | 6 (18.2%) | 0 (0.0%) |
| 3 | 1 (2.8%) | 0 (0.0%) | 6 (20.0%) |
| ≥4 | 2 (5.6%) | 0 (0.0%) | 18 (60.0%) |
| Episode length[c] | | | |
| Min; Max | 94 min; 10,959 min (=7.5 d) | 19,298 min; 42,600 min (=29.5 d) | 5 min; 751 min |
| Median | 1349 min | 22,146 min | 46 min |
| Interval between episodes[c] | | | |
| Min; Max | 1 min; 43,290 min (=30 d) | 1440 min; 106,821 min (=74 d) | 1 min; 8681 min (=6 d) |
| Median | 2267 min (=1.5 d) | 2281 min (=1.5 d) | 315 min |
| Test setting | | | |
| Smallest available frequency of testing[d] | Per minute (time of episode's start and end given in format hh:mm) | Per minute (time of episode's start and end given in format hh:mm) | Per minute (time of episode's start and end given in format hh:mm) |
| Index test | Rule-based detection model | Not available (proxy: altered motor activity) | Not available (proxy: altered electroencephalogram recording) |
| Reference standard | Clinician's diagnosis | Not available (proxy: motor activity) | Electroencephalogram recording |
| Ideal estimation set(s) for diagnostic test accuracy | | | |
| Estimation level(s) | Time-level and/or block-level with blocks based on reference standard | | |
| Time unit(s) | Per minute and/or hour | Per minute, hour, and/or day | Per minute |

[a] SIRS = Systematic Inflammatory Response syndrome.

[b] Patient with two records and the patient, who was later added, are added here due to lack of information on gender as well as the additional six disease-free cases.

[c] Only on participants who experience at least one episode based on the reference standard.

[d] The smallest available frequency of testing allows to summarise the testing using any other larger testing frequency (ie, per minute can be transformed to, eg, per hour, per day, …).

center in Germany between August 01, 2018 and March 31, 2019. For details, we refer to [23–25]. We used the data of 36 consecutive patients (10 disease-free) to ensure comparability with the other datasets regarding the sample size and to comply with the restrictions imposed by the data owners.

### 2.3.2. *Depression dataset*

The depression dataset includes records of 55 adult patients, of whom 23 experienced a depressive episode. All individuals were recruited at Haukeland University, Norway. In total, data from 693 days was recorded [26]. For this study's purpose, the dataset included all cases and only the first 10 disease-free controls ($n = 33$ patients).

### 2.3.3. *Epilepsy dataset*

The epilepsy dataset entails electroencephalogram (EEG) recordings of 24 pediatric and young adult patients with intractable seizures of the Boston Children's Hospital, in the USA. Each patient was likely to develop an epileptic episode due to having stopped the antiseizure medication under medical supervision in an inpatient setting. A total of 197 episodes were recorded. Modifications were applied to the dataset to meet this study's research purpose: Six additional disease-free synthetic patient records were added (ie, total sample size: 30 patients).

## 2.4. *Analysis*

Sensitivities and specificities were estimated for each diagnosis per time unit (ie, minute, hour, and day) and per estimation level (ie, time-level, block-level, and patient-time-level) using the labeling as shown above (Appendix 2–3 and Fig 2) and the nonparametric approach [11,19]. For comparability, sensitivities and specificities were also estimated ignoring data clustering (ie, using the standard formulae (Table 1) with a) time-level data [time unit: minute] or b) only one label per patient [''patient-level'': TP or FN for a diseased patient; TN or FP for a disease-free patient]). Missing values were not observed. Indeterminate test results were not registered.

## 3. Results

We observed relevant differences across and within the use cases for the three estimation levels and time units (Fig 4 and Table 3).

Across the use cases, we observed that the highest DTAs, irrespective of the used time unit and/or diagnosis, were estimated on the time-level, while the DTAs of the block-levels and patient-time-level were lower. The block-level analysis with blocks based on the reference standard showed that the unmodified DTA estimates were lower than the DTA estimates after index test correction. Moreover, the DTA estimates using 'day' as a time unit were closer to 100%, irrespective of the estimation level, than the DTA estimates using 'minute' or 'hour' as a time unit. We observed a similar pattern of DTA estimates for the time units 'minute' and 'hour' across the individual estimation levels, although estimates were slightly higher given the 'hour' time unit. An exception was the sensitivity of the use case of depression which was identical for the individual estimation levels irrespective of the used time unit. Furthermore, the number of observations decreased dramatically from the time-level to the block-level and/or patient-time-level which is somewhat mirrored by the estimates and their CIs. The comparison of the DTA evaluation accounting for data clustering vs ignoring data clustering highlighted the risk of distorted estimates. If all test results were included in the estimation but considered uncorrelated, the point estimates were identical to the time-level, but their corresponding CIs were narrower. If, however, all test results were aggregated into a single test result per patient, the sensitivities were higher and the specificities lower compared to the time-level, while the CIs were wider.

The within-use-case comparison showed that some of the DTA evaluations were clinically more meaningful and informative than others considering the disease-specific characteristics (Tables 2–3). For SIRS, the time units 'minute' and/or 'hour' were most relevant considering the index test. The sensitivities and specificities for these time units were relatively similar (ie, a maximum difference of 10.8 percentage points), and dataset-specific information on the episode and disease-free interval lengths highlighted that differences between the diagnostic tests would be undetected if the time unit 'day' were to be used. Generally, the DTA estimates were higher using the time unit 'day' than any of the smaller time units except for the specificities of the time unit 'day' on the time-level (94.4%) and the modified block-level with blocks based on the reference standard (90.5%). For depression, we observed that the larger the time unit, the smaller the differences between the point estimates across the estimation levels. The block-level with blocks based on the reference standard showed higher DTAs after correcting the index test (ie, increases of 5 percentage points for sensitivity and 30–40 percentage points for specificity). Finally, for epilepsy, we observed that the time unit 'minute' was the most meaningful considering the index test because the length of episodes and disease-free intervals were on a smaller scale. The DTA estimates using the time unit 'hour' were higher than the estimates of the time unit 'minute', which were even further elevated for the time unit 'day'.

## 4. Discussion

Our study shows that two features – estimation level and time unit – should be considered in accordance with diagnosis-specific characteristics so that the test is

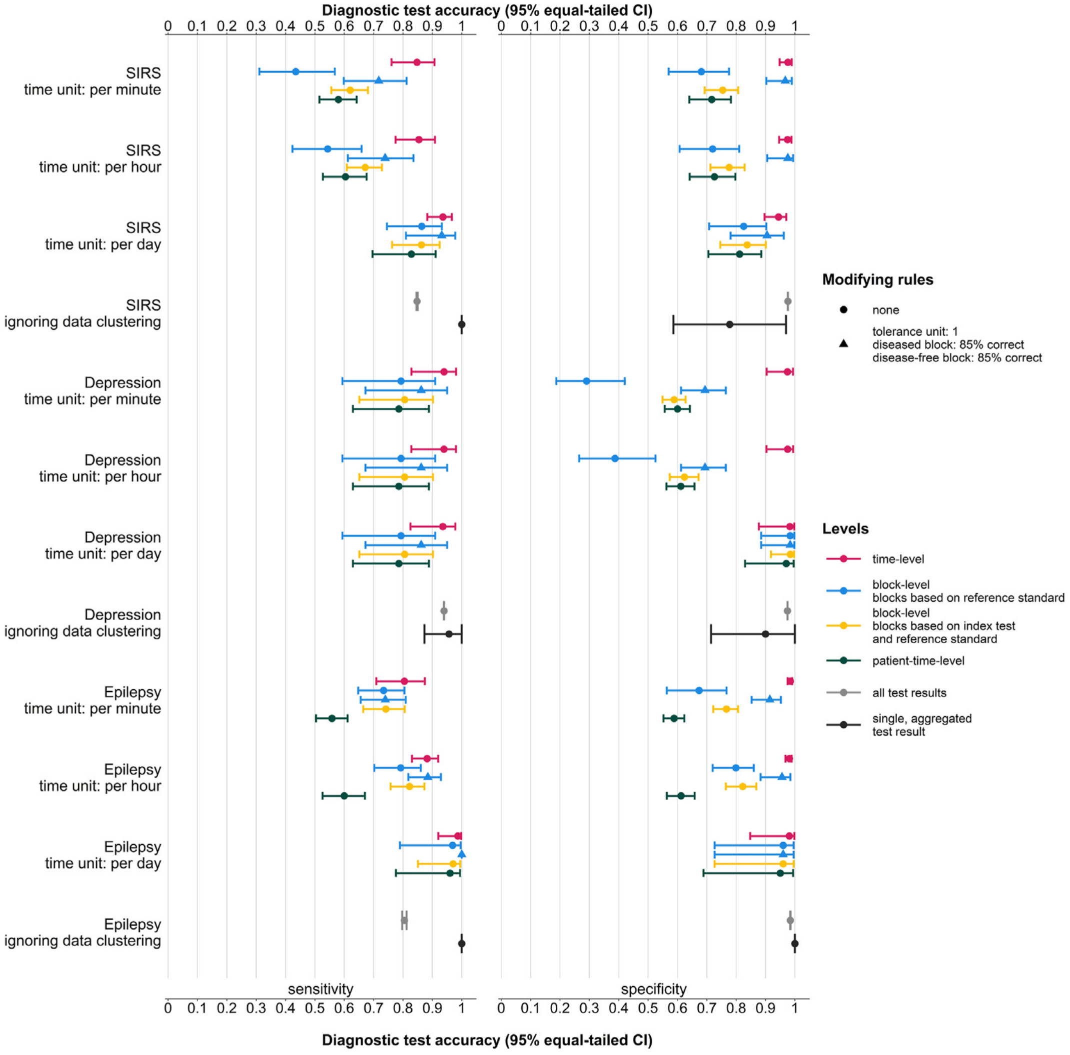


**Figure 4.** Summary of the diagnostic test accuracy of all three diagnoses stratified by the diagnostic test accuracy indices (ie, sensitivity and specificity), by the estimation level (i.e, time-level, block-level, and patient-time-level), and by the time unit (ie, minute, hour, and day). The levels 'all test results' and 'single, aggregated test result' (aka patient-level) present the diagnostic test accuracy when data clustering is ignored. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

meaningful and decision-informing (Table 4) when presenting an index test's DTA in a longitudinal setting. They are determining the number of observations included in the DTA estimation. Estimand and statistical approach should be chosen appropriately, that is, accounting for data clustering [18]. Using a simple approach without accounting for the data structure leads to an overestimation of DTA and its precision, especially if using a single aggregated test result, when compared with what is relevant for clinical practice.

Various estimation levels can be used for the analysis and reporting of an index test'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. So far, most studies have reported their analytical procedures and reporting level [27] intransparently; only few provided details on the estimation level. As an example, Wulff et al. [23] used the time- and patient-time-level. Bode et al. [28] used the block-level with blocks based on both diagnostic tests. The time-level includes every single time point which makes it a good technical starting point, while the block-level summarizes these time points per diseased and disease-free block into a single, block-specific label. This requires that the blocks are based on the diagnostic status of the reference standard so that the length of the blocks is fixed. We recommend using the time-level when having a disease with short episodes (eg, epilepsy) or to assess the index test's precision (ie, maximum of labels defined by time unit). The block-level with blocks based on reference standard can be used if the aim is to assess the index test's performance in a clinical setting (ie, focus on periods that have correctly or incorrectly been classified by the index test) without having a constant decision to make. For example, a rule-based detection model could be used in the intensive care setting

**Table 3.** Summary of the diagnostic test accuracy per diagnostic level (ie, per time-level, per block-level, and per patient-time-level) per time unit (ie, per minute, per hour, and per day) for the diagnoses 'Systemic Inflammatory Response Syndrome' (SIRS), depression, and epilepsy. The levels 'all test results' and 'single, aggregated test result' (aka patient-level) present the diagnostic test accuracy when data clustering is ignored

| 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 |
| | | Block[a] *(blocks based on reference standard)* | 43.5% (31.1–56.8%) | 68.1% (57.0–77.5%) | 20 | 29 | 26 | 62 |
| | | Block[b] *(blocks based on reference standard)* | 71.7% (59.8–81.2%) | 96.7% (90.2–98.9%) | 33 | 3 | 19 | 88 |
| | | Block *(blocks based on index test and reference standard)* | 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%) | 1454 | 216 | 248 | 8315 |
| | | Block[a] *(blocks based on reference standard)* | 54.3% (42.3–65.9%) | 72.0% (60.7–81.0%) | 25 | 23 | 21 | 59 |
| | | Block[b] *(blocks based on reference standard)* | 73.9% (61.2–83.6%) | 97.6% (90.6–99.4%) | 34 | 2 | 12 | 80 |
| | | Block *(blocks based on index test and reference standard)* | 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 |
| | | Block[a] *(blocks based on reference standard)* | 86.4% (74.5–93.2%) | 82.5% (70.7–90.2%) | 38 | 11 | 6 | 52 |
| | | Block[b] *(blocks based on reference standard)* | 93.2% (81.0–97.8%) | 90.5% (78.0–96.2%) | 41 | 6 | 3 | 57 |
| | | Block *(blocks based on index test and reference standard)* | 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 |
| | Not applicable | All test results *(ignoring data clustering)* | 84.8% (84.6–85.0%) | 97.6% (97.6–97.6%) | 84,243 | 12,315 | 15,124 | 500,847 |
| | Not applicable | Single, aggregated test result *(ignoring data clustering)* | 100% (100–100%) | 77.8% (58.6–97.0%) | 30 | 4 | 0 | 14 |
| Depression | Minute | Time | 93.9% (82.9–98.0%) | 97.5% (90.4–99.4%) | 728,122 | 75,338 | 46,926 | 2,886 594 |
| | | Block[a] *(blocks based on reference standard)* | 79.3% (59.4–91.0%) | 29.0% (18.7–42.0%) | 23 | 44 | 6 | 18 |
| | | Block[b] *(blocks based on reference standard)* | 86.2% (67.2–95.0%) | 69.4% (61.2–76.4%) | 25 | 19 | 4 | 43 |
| | | Block *(blocks based on index test and reference standard)* | 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 |

(*Continued*)

**Table 3.** Continued

| Diagnosis | Time unit | Level | Sensitivity (95% CI) | Specificity (95% CI) | True Positive | False Positive | False negative | True Negative |
|---|---|---|---|---|---|---|---|---|
| | Hour | Time | 93.9% (82.8–98.0%) | 97.5% (90.3–99.4%) | 12,159 | 1232 | 788 | 48,104 |
| | | Block[a] *(blocks based on reference standard)* | 79.3% (59.4–91.0%) | 38.7% (26.5–52.5%) | 23 | 38 | 6 | 24 |
| | | Block[b] *(blocks based on reference standard)* | 86.2% (67.2–95.0%) | 69.4% (61.2–76.4%) | 25 | 19 | 4 | 43 |
| | | Block *(blocks based on index test and reference standard)* | 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 | 2045 |
| | | Block[a] *(blocks based on reference standard)* | 79.3% (59.4–91.0%) | 98.4% (88.5–99.8%) | 23 | 1 | 6 | 61 |
| | | Block[b] *(blocks based on reference standard)* | 86.2% (67.2–95.0%) | 98.4% (88.5–99.8%) | 25 | 1 | 4 | 61 |
| | | Block *(blocks based on index test and reference standard)* | 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 |
| | Not applicable | All test results *(ignoring data clustering)* | 93.9% (93.9–94.0%) | 97.5% (97.4–97.5%) | 728,122 | 75,338 | 46,926 | 2,886,594 |
| | Not applicable | Single, aggregated test result *(ignoring data clustering)* | 95.7% (87.3–100%) | 90.0% (71.4–100%) | 22 | 1 | 1 | 9 |
| Epilepsy | Minute | Time | 80.5% (70.9–87.4%) | 98.5% (97.5–99.0%) | 8485 | 2409 | 2061 | 153,075 |
| | | Block[a] *(blocks based on reference standard)* | 73.4% (64.8–80.5%) | 67.3% (56.4–76.7%) | 124 | 65 | 45 | 134 |
| | | Block[b] *(blocks based on reference standard)* | 74.0% (65.6–80.9%) | 91.5% (85.2–95.2%) | 125 | 17 | 44 | 182 |
| | | Block *(blocks based on index test and reference standard)* | 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 | 2422 |
| | | Block[a] *(blocks based on reference standard)* | 79.2% (70.2–86.1%) | 79.9% (72.0–86.1%) | 103 | 32 | 27 | 127 |
| | | Block[b] *(blocks based on reference standard)* | 88.5% (81.8–92.9%) | 95.6% (88.3–98.4%) | 115 | 7 | 15 | 152 |
| | | Block *(blocks based on index test and reference standard)* | 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 |
| | | Block[a] *(blocks based on reference* | 96.9% (79.0–99.6%) | 96.0% (72.6–99.5%) | 31 | 1 | 1 | 24 |

(*Continued*)

**Table 3.** Continued

| Diagnosis | Time unit | Level | Sensitivity (95% CI) | Specificity (95% CI) | True Positive | False Positive | False negative | True Negative |
|---|---|---|---|---|---|---|---|---|
| | | *standard)* | | | | | | |
| | | Block[b] *(blocks based on reference standard)* | 100% (100–100%) | 96.0% (72.6–99.5%) | 32 | 1 | 0 | 24 |
| | | Block *(blocks based on index test and reference standard)* | 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 |
| | Not applicable | All test results *(ignoring data clustering)* | 80.5% (79.7–81.2%) | 98.5% (98.4–98.5%) | 8485 | 249 | 2064 | 153,075 |
| | Not applicable | Single, aggregated test result *(ignoring data clustering)* | 100% (100–100%) | 100% (100–100%) | 24 | 0 | 0 | 6 |

[a] Block-level based on reference standard: No correction rules applied.

[b] Block-level based on reference standard: Application of ±1 time point tolerance margin at start/end of reference standard episode and 85%-correction within diseased and disease-free blocks.

to help clinicians detecting SIRS episodes [23,29] so that they are alerted when the patient's health deteriorates.

Diagnosis-specific characteristics must be considered before performing the DTA estimation, as they determine the required time unit. Many DTA studies withhold sufficient information on their time unit, how they account for the inflation of the type 1 error in the DTA estimation [30], and/or whether they used longitudinal data [8]. We identified few studies (eg, [8,23,28,31–35]) that indicated/hinted at their used time unit. As with the estimation level, the used time unit influences the interpretation and understanding of the DTA estimates [30]. An inappropriate time unit, especially if too large, causes an increase in TP and TN observations so that the estimates are distorted due to losing information about diagnostic test differences. For example, epileptic seizures last seconds to minutes, which excludes 'day' as a time unit. In our use case, the DTA estimates using 'day' as a time unit approached 100%, showing barely any differences between the estimation levels. However, a smaller unit increases accuracy [36], which we could not determine to be potentially misleading. For example, given diagnosis-specific characteristics of depression, we considered 'day' to be an adequate time unit, but with the smaller time units, we observed differences between the diagnostic tests potentially driven by episode onset and end classifications; thus, they are not incorrect. Diseases characterized by medium to long episode periods and disease-free intervals between episodes, such as SIRS [37] or depression [38–40], can be assessed using any of the three time units. Moreover, the date-time classification of an episode should be specific. If, eg, both tests classify per day (ie, starting at 00:00 AM and finishing at 11:59 PM), then the DTA estimates are identical irrespective of time unit and estimation level. This is caused by equally inflating the number of observations included in the clusters in comparison to fewer numbers of observations. We suggest using a time unit that best represents the diagnosis-specific characteristics. If in doubt, smaller time units are preferable because they allow for the precise assessment the episode start and end date-times. However, the translation of observed DTA into clinically meaningful DTA is often hampered 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; thus, they are subject to a certain risk of data-generating pitfalls [41]. Especially the depression and epilepsy datasets lacked information on index tests and reference standard diagnoses; hence, index tests and reference standard diagnoses were produced based on the available information in the datasets. Incorporation bias is most likely present in both datasets [42]. However, for this study's purpose, it remains

**Table 4.** Generalized interpretation of diagnostic test accuracy estimates per estimation level under the consideration of the time unit. These should be a reflection of the research question's estimand(s)

| Level | Interpretation |
|---|---|
| Time | Throughout the whole observation period, the index test detects at each time point (here: per *[time unit]*) the presence of *[diagnosis]* with a sensitivity of *[sensitivity]*% (95% CI: *lower-upper*%) and the absence of *[diagnosis]* with a specificity of *[specificity]*% (95% CI: *lower-upper*%) given the ground truth definition of the reference standard. |
| Block[a] *(based on reference standard)* | Throughout the whole observation period, the index test detects *[diagnosis]* events (ie, period of a *[diagnosis]* episode) with a sensitivity of *[sensitivity]*% (95% CI: *lower-upper*%) and disease-free periods with a specificity of *[specificity]*% (95% CI: *lower-upper*%) given the ground truth definition of the reference standard. |
| Block[b] *(based on reference standard)* | Throughout the whole observation period, the index test detects *[diagnosis]* events (ie, period of a *[diagnosis]* episode) with a sensitivity of *[sensitivity]*% (95% CI: *lower-upper*%) and disease-free periods with a specificity of *[specificity]*% (95% CI: *lower-upper*%) given the ground truth definition of the reference standard after allowing the episode start and end to deviate by ± *[tolerance margin] [time unit]* and correcting of individual incorrect time points (ie, per *[time unit]*) per block if $\geq$ *[percent correctness]*% of the block-specific time points were correctly classified. |

[a] Block-level based on reference standard: No correction rules applied.

[b] Block-level based on reference standard: Application of ± $t$ time point tolerance margin at start/end of reference standard episode and/or $p$ %-correction within diseased and disease-free blocks.

unconcerning because we aimed to demonstrate the problem of estimating an index test's DTA using longitudinal data.

In this study, we assumed that the reference standards perfectly diagnosed the diseases. Depending on the clinical setting, this might not be true, especially in situations where the diagnostic test is expected to alert clinicians before the reference standard becomes positive. Researchers should keep in mind that the index tests and/or reference standard can change over time (eg, updated guidelines for diagnosis).

## 5. Conclusion

Using longitudinal data in a DTA study requires researchers to consider methodological choices and a clear, predefined estimand early in the planning phase. Choices need to be made on the estimation level(s) and the time unit(s) considering diagnostic-specific characteristics. When reporting the DTA study's findings, researchers should be transparent and state their rationales. 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 (eg, missing values).

## Ethics approval

We used only publicly available anonymized datasets so no approval by an institutional review board was required.

## 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.

## Data statement

The original datasets can be accessed via the data owners (see "2.3 The datasets"); the modified-labelled 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-labelled datasets remain with the data owners of this publication. The R package "diagacc" can be accessed via https://github.com/wbr-p/diagacc. All analyses were conducted using R version 4.2.3 (2023-03-15).

## CRediT authorship contribution statement

**Julia Böhnke:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Visualization, Writing – original draft, Writing – review & editing. **Antonia Zapf:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Supervision, Visualization, Writing – original draft, Writing – review & editing. **Katharina Kramer:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Supervision, Visualization, Writing – original draft, Writing – review & editing. **Philipp Weber:** Formal analysis, Writing – original draft, Writing – review & editing. **Louisa Bode:**

Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Supervision, Visualization, Writing – original draft, Writing – review & editing. **André Karch:** Writing – original draft, Writing – review & editing. **Nicole Rübsamen:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Supervision, Visualization, Writing – original draft, Writing – review & editing.

## Data availability

Data will be made available on request.

## Declaration of Competing interest

The authors declare that they have no competing interests.

## Acknowledgments

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.

## Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jclinepi.2024.111314.

## References

[1] Definition of diagnostic test - NCI dictionary of cancer terms - national cancer institute n.d. Available at: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/diagnostic-test. Accessed March 31, 2022.

[2] Leeflang MMG, Allerberger F. How to: evaluate a diagnostic test. Clin Microbiol Infect 2019;25:54–9.

[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/EEC; 2017. Available at: http://data.europa.eu/eli/reg/2017/745/oj. Accessed February 2, 2022.

[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;118:550–60.

[5] Chassé M, Fergusson DA. Diagnostic accuracy studies. Semin Nucl Med 2019;49:87–93.

[6] Sitch AJ, Dekkers OM, Scholefield BR, Takwoingi Y. Introduction to diagnostic test accuracy studies. Eur J Endocrinol 2021;184:E5–9.

[7] Miller DC, Dunn RL, Wei JT. Assessing the performance and validity of diagnostic tests and screening programs. In: Penson DF, Wei JT, editors. Clinical research methods for surgeons. Totowa, NJ: Humana Press; 2007:157–74. https://doi.org/10.1007/978-1-59745-230-4_10.

[8] Böhnke J, Varghese J, Karch A, Rübsamen N, Bode L, Mast M, et al. Systematic review identifies deficiencies in reporting of diagnostic test accuracy among clinical decision support systems. J Clin Epidemiol 2022;151:171–84.

[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;267:581–8.

[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;265:910–6.

[11] Lange K. Nichtparametrische Analyse diagnostischer Gütemaße bei Clusterdaten. Germany: Georg-August-University Göttingen; 2011.

[12] Gönen M, Panageas KS, Larson SM. Statistical issues in analysis of diagnostic imaging experiments with multiple observations per patient. Radiology 2001;221:763–7.

[13] Mondol MH, Rahman MS. Bias-reduced and separation-proof GEE with small or sparse longitudinal binary data. Stat Med 2019;38:2544–60.

[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;5:169–88.

[15] Committee for Medicinal Products for Human Use of the 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. Available at: https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e9-r1-addendum-estimands-and-sensitivity-analysis-clinical-trials-guideline-statistical-principles-clinical-trials-step-5_en.pdf. Accessed May 12, 2021.

[16] Lawrance R, Degtyarev E, Griffiths P, Trask P, Lau H, D'Alessio D, 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 Rep Outcomes 2020;4:68.

[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 Ärztebl Int 2021;118:883.

[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). Düsseldorf: German Medical Science GMS Publishing House; 2022: The estimand framework in diagnostic studies.

[19] Konietschke F, Brunner E. Nonparametric analysis of clustered data in diagnostic trials: estimation problems in small sample sizes. Comput Stat Data Anal 2009;53:730–41.

[20] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. Stard 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015;351:h5527.

[21] Cui Y, Konietschke F, Harrar SW. The nonparametric Behrens-Fisher problem in partially complete clustered data. Biom J 2021;63:148–67.

[22] Lange K, Brunner E. Sensitivity, specificity and ROC-curves in multiple reader diagnostic trials—a unified, nonparametric approach. Stat Methodol 2012;9:490–500.

[23] Wulff A, Montag S, Rübsamen N, Dziuba F, Marschollek M, Beerbaum P, 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 Inf Decis Making 2021;21:62.

[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;295:100–3.

[25] Wulff A, Mast M, Bode L, Marschollek M, Schamer S, Beerbaum P, et al. ELISE - An open pediatric intensive care data set 2022. Available at: https://leopard.tu-braunschweig.de/receive/dbbs_mods_00070468. Accessed August 25, 2022.

[26] Garcia-Ceja E, Riegler M, Jakobsen P, Tørresen J, Nordgreen T, Oedegaard KJ, et al. Depresjon: A motor activity database of depression episodes in unipolar and bipolar patients. Proceedings of the 9th ACM Multimedia Systems Conference, MMSys 2018; 2018:472−7. Available at: https://dl.acm.org/doi/10.1145/3204949.3208125. Accessed January 10, 2022.

[27] Westwood M, Joore M, Grutters J, Redekop K, Armstrong N, Lee K, 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 2013;17:7−243.

[28] Bode L, Schamer S, Böhnke J, Study Group E, Bott O, Marschollek M, et al. Tracing the progression of sepsis in critically ill children: clinical decision support for detection of hematologic dysfunction. Appl Clin Inf 2022;13:1002−14.

[29] Mast M, Marschollek M, Jack T, Wulff A, Elise Study Group. Developing a data driven approach for early detection of SIRS in pediatric intensive care using automatically labeled training data. Stud Health Technol Inform 2022;289:228−31.

[30] Parsons NR, Teare MD, Sitch AJ. Unit of analysis issues in laboratory-based research. Elife 2018;7:e32486.

[31] Dewan M, Muthu N, Shelov E, Bonafide CP, Brady P, Davis D, et al. Performance of a clinical decision support tool to identify PICU patients at high risk for clinical deterioration. Pediatr Crit Care Med 2020;21:129−35.

[32] Nagori A, Dhingra LS, Bhatnagar A, Lodha R, Sethi T. Predicting hemodynamic shock from thermal images using machine learning. Sci Rep 2019;9:91.

[33] Calvert JS, Price DA, Chettipally UK, Barton CW, Feldman MD, Hoffman JL, et al. A computational approach to early sepsis detection. Comput Biol Med 2016;74:69−73.

[34] 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.

[35] Wulff A, Montag S, Steiner B, Marschollek M, Beerbaum P, Karch A, 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:e028953.

[36] Hess AS, Shardell M, Johnson JK, Thom KA, Strassle P, Netzer G, et al. Methods and recommendations for evaluating and reporting a new diagnostic test. Eur J Clin Microbiol Infect Dis 2012;31.

[37] Chakraborty RK, Burns B. Systemic inflammatory response syndrome. StatPearls; 2022. Available at: https://www.ncbi.nlm.nih.gov/books/NBK547669/. Accessed August 17, 2022.

[38] Chand SP, Arif H. Depression. StatPearls. Treasure Island, FL: StatPearls Publishing; 2022. Available at: https://www.ncbi.nlm.nih.gov/books/NBK430847/. Accessed November 15, 2022.

[39] Goodwin G. Depression. In: Castle D, Coghill D, editors. Comprehensive Men's Mental Health. Cambridge: Cambridge University Press; 2021:128−38.

[40] Strunk DR, Pfeifer BJ, Ezawa ID. Depression. In: Wenzel A, editor. Handbook of cognitive behavioural therapy: Applications, Vol. 2. Washington, DC, US: American Psychological Association; 2021:3−31.

[41] Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med 2019;38:2074−102.

[42] Catalogue of Bias Collaboration. Incorporation bias. In: Plüddemann A, McCall M, editors. Scakett Catalogue of Biases; 2019. Available at: https://catalogofbias.org/biases/incorporation-bias/. Accessed May 25, 2022.