Additional file 5: Statistical analysis plan

This statistical analysis plan (SAP) is specific to the *data analysis* of the ELISE study. A statistical analysis is not intended in either the *enrolment and clinical assessment* or the *model assessments*; hence, any interim analyses are strictly prohibited.

1. Outcomes

1.1 Primary confirmatory analyses

In total, five primary confirmatory analyses are conducted to determine the performance of the predictive Clinical Decision Support System (CDSS) algorithms for (1) SIRS (here: main outcome), (2) sepsis, (3) hepatic organ dysfunction (OD), (4) hematologic OD, and (5) respiratory OD. These hypotheses are tested hierarchically (i.e., not a multi-diagnostic approach) in the order as listed; hence, the type I error does not need to be adjusted. Each analysis estimates the sensitivity and specificity (co-primary endpoints) of the predictive model for the specific target condition. It will be tested whether the predictive model has a sensitivity ≥ 75% and a specificity ≥ 75% in detecting the target condition and whether the lower limits of the 95% Wald Confidence Interval (CI) of sensitivity and specificity are both greater than 75%. Only if this is the case, the next hypothesis is tested in the mentioned hierarchical order. The threshold of 75% marks the minimum DTA with which our prediction models should diagnose to still be considered a useful support for clinicians.

1.2 Secondary exploratory analyses

Two secondary exploratory analyses are conducted to determine the performance of the predictive models of renal OD and cardiovascular OD due to their rather small number of expected cases. The analysis is identical to the approach of primary confirmatory analyses and the results could be used as preparation for sample size planning for future DTA studies for these endpoints.

Moreover, an additional secondary exploratory analysis is conducted that compares the sensitivities and specificities of the predictive models for SIRS, sepsis and associated ODs (one at a time) with the corresponding routine evaluations of the clinicians in the respective shift before the onset (up to 12 hours) of the disease. The comparison of two diagnostic approaches (predictive CDSSs vs standard of care) per target condition, and if the predictive CDSSs are

superior to the clinician assessments, is evaluated with the McNemar χ^2 test [1]. These results are interpreted only in an explorative approach.

2. Labelling

The predictive CDSS estimates the probability of the endpoint every x^{th} minute (where x is a constant that may be different dependent on the endpoint) based on the data acquired before the time of estimation. This way, a patient's paediatric intensive care unit (PICU) stay is divided into a sequence of measurement units of x minutes. Per patient and PICU stay, each measurement unit is labelled as either True Positive (TP), False Positive (FP), False Negative (FN), or True Negative (TN) as explained in Table 1.

TP	A measurement unit in which both the predictive CDSS and the reference standard detect the occurrence of the target disease.
FP	A measurement unit in which the predictive CDSS detects the occurrence of the target disease but the reference standard does not.
FN	A measurement unit in which the reference standard detects the occurrence of the target disease but the predictive CDSS does not.
TN	A measurement unit in which both the predictive CDSS and the reference standard do not detect the occurrence of the target disease.

Table 1: Possible classification labels per measurement unit when comparing two of the diagnostic approaches.

The resulting labels are then slightly modified by applying the following *mandatory rules*:

- *Merging*: Diagnostic episodes (i.e., periods where the reference standard or the CDSS are positive) with less than y hours in-between the end of one diagnostic episode and the start of a new diagnostic episode are merged to one episode. To our knowledge, information to clearly distinguish episodes are not defined yet; hence, this will be decided on an individual basis per diagnosis. However, episodes with more than 48 hours in-between are considered to be independent episodes within this study.
- *Death*: The death of a patient is equal to the end of the last episode and the discharge date of that particular patient.

- *PICU stay*: Diagnostic episodes before or after the paediatric intensive care unit (PICU) stay are excluded from the diagnostic test accuracy (DTA) estimation; thus, they will not be labelled.

Additionally, the predictive CDSS models may not be precise enough to meet the exact timepoint of the start and/or end of the gold standard episode although in practice the models should be pretty close. Based on the data, the following *optional rule* may be used:

- *±z-hour window*: The index test's diagnostic episode start and end is TP if the index test's diagnostic episode start or end happened within the ±z-hour window around the reference standard diagnostic episode start or end. Hence, FN or FP measurement units within this time window are converted to TP.

3. Statistical methods

A clustered nonparametric approach [2–4] is used to estimate sensitivities and specificities with their 95% Wald CI for the above-mentioned outcomes (i.e., no p-value evaluation). This method is used because it accounts for the longitudinal data format (i.e., data per day and several days per patient) by weighting each cluster (here: a single subject) according to the number of observations (measurement units) per cluster; thus, in the weighted estimates each observation has the same weight while in unweighted estimates each patient is given the same weight. Additionally, the estimations are adjusted for the imperfect gold standard [5], the previously evaluated knowledge-based models, from which the sensitivities and specificity are known from previous evaluation processes [6]. The imperfect gold standard is used because it is time-efficient to view and assess all data per patient rather than having clinicians to perform the viewing and assessment of large amounts of patient data. This approach is applied to all primary confirmatory outcomes and the first and second secondary exploratory outcomes (i.e., renal OD and cardiovascular OD).

The third exploratory outcome is a comparison of the sensitivities and specificities of two diagnostic approaches (i.e., predictive models vs standard of care) using the McNemar χ^2 test. Here the results are interpreted given a significance level of $\alpha = 0.05$.

4. Subgroup analyses

In addition to the main analysis, three subgroup analyses are planned to test the performance of the each predictive CDSS models for various patient strata to determine whether the predictive model under investigation performs equally well regardless of age and/or sex. Those subgroup analyses are performed because the diagnose-specific models are uniformly applied to all participants; thus, it may not perform equally well for all subgroups. All subgroup analyses test the previously listed outcomes given the above-mentioned statistical tests. The subgroup analyses are:

1) Stratification by age groups[1]
2) Stratification by sex[2]
3) Stratification by age groups[1] and sex[2]

Analyses may not be possible if none or too few individuals per strata could be recruited. All subgroup analyses are exploratory as the power for testing decreased due to the reduced sample size.

5. Analysis population

Included in the final analysis will be all patients meeting the eligibility criteria (PICU stay ≥ 12 hours and valid informed consent). Individuals who previously provided their consent but withdraw their consent before the *model assessments* will not be part of the final analysis. Their data is neither assessed by the predictive CDSS models nor by the knowledge-based CDSS models. Should the participation be withdrawn during the *model assessments*, these individuals will also not be included in the final analysis since the final analysis is performed in the *data analysis*. A withdrawal during the *data analysis* is not possible since the data is anonymised and already assessed by the three assessors. Nonetheless, reasons for withdrawal in the various phases will be reported.

[1] **Age groups:** NEWBORNS (0 days to 1 week = 0 to 7 days), NEONATALES (1 week to 1 month = 8 to 30 days), INFANTS (1 month to 1 year = 31 to 365 days), TODDLERS and PRE-SCHOOL CHILDREN (2 to 5 years = 366 to 1,825 days), SCHOOL CHILDREN (6 to 12 years = 1,826 to 4,380 days), and ADOLESCENTS (13 to <18 years = 4,381 to 6,443 days)

[2] **Sex:** MALE and FEMALE

6. Handling missing data

This is a DTA study evaluation using a longitudinal data format; thus, the total PICU stay of the patient will be assessed by the diagnostic test and are afterwards subject to labelling. In case of missing values, the predictive CDSS models and the knowledge-based CDSS models carry the last observation forward.

Missing values in the clinician assessment will be addressed by only including observations in the analysis for which the clinicians documented a disease assessment (i.e., *complete case analysis*). Additionally, we will quantify and report the impact of missing clinician assessments by estimating the sensitivities and specificities by assuming:

- *Best-Case-Scenario*: All missing disease assessments by clinicians are rated correctly (i.e., the missing disease assessment are imputed as either TP or TN); *OR*
- *Worst-Case-Scenario*: All missing disease assessments by clinicians are rated incorrectly (i.e., the missing disease assessments are imputed as either FP or FN).

7. Software

The labelling and all analyses will be conducted using the latest available R version. The R code will be made publicly available at GitLab upon publication of the results.

8. References

1 Trajman A, Luiz RR. McNemar chi2 test revisited: comparing sensitivity and specificity of diagnostic examinations. *Scand J Clin Lab Invest* 2008;**68**:77–80. doi:10.1080/00365510701666031

2 Brunner E, Zapf A. Nonparametric ROC Analysis for Diagnostic Trials. *Methods Appl Stat Clin Trials* 2014;**2**:483–95. doi:10.1002/9781118596333.CH28

3 Lange K. *Nichtparametrische Analyse diagnostischer Gütemaße bei Clusterdaten*. [Dissertation]: University of Goettingen, 2011. Accessed January 20, 2022. https://ediss.uni-goettingen.de/bitstream/handle/11858/00-1735-0000-000D-F1D1-B/lange.pdf?sequence=1

4 Rooney D. *Covariate adjusted nonparametric estimation of sensitivity and specificity*

in clustered data. [Master thesis]: Heidelberg University Hospital, 2017.

5 Umemneku Chikere CM, Wilson KJ, Allen AJ, *et al.* Comparative diagnostic accuracy studies with an imperfect reference standard – a comparison of correction methods. *BMC Med Res Methodol* 2021;**21**:1–12. doi:10.1186/S12874-021-01255-4/TABLES/8

6 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**:1–9. doi:10.1186/S12911-021-01428-7/TABLES/3