Additional file 2: Sample size

The hypotheses for the sample size calculation were defined as follows:

1. For the predictive model for Systemic Inflammatory Response Syndrome (SIRS) detection, the null hypothesis is H_0: $\{Sensitivity\ <\ 75\%\ \cup\ Specificity\ <\ 75\%\}$ (intersection union test).
2. For the predictive model for sepsis, the null hypothesis is H_0: $\{Sensitivity\ <\ 75\%\ \cup\ Specificity\ <\ 75\%\}$ (intersection union test).
3. For the predictive model for hepatic organ dysfunction, the null hypothesis is H_0: $\{Sensitivity\ <\ 75\%\ \cup\ Specificity\ <\ 75\%\}$ (intersection union test).
4. For the predictive model for haematological organ dysfunction, the null hypothesis is H_0: $\{Sensitivity\ <\ 75\%\ \cup\ Specificity\ <\ 75\%\}$ (intersection union test).
5. For the predictive model for respiratory organ dysfunction, the null hypothesis is H_0: $\{Sensitivity\ <\ 75\%\ \cup\ Specificity\ <\ 75\%\}$ (intersection union test).

Hypotheses 1 to 5 are tested hierarchically in the presented order. Only if all previous hypotheses could be rejected, the next hypothesis can be tested for confirmation.

The sample size was calculated using the method of Stark and Zapf [1]. This approach ensures the desired overall power which is perfectly adjusted to the prevalence. The sample size is optimal in the way that it is the smallest representative sample that achieves the advertised overall power. The approach is based on the idea to individually split the overall power to the endpoint of the sensitivity and specificity. Hence, an individual type II error is assigned to each of both endpoints so that the required sample sizes of both endpoints are equal. To reach an overall power of 80%, the individual power of each endpoint cannot be smaller than 80%.

The pre-test probability (prevalence) of SIRS (primary endpoint, hypothesis 1) is expected to be 60% (data from the CADDIE-2 study [2]. The estimators for the diagnostic accuracy of the predictive model were 85% for sensitivity and 84% for specificity. For the sample size calculation, a one-sided type I error of 2.5% was assumed and the global power was set to 80% (as described above). For a sensitivity of 75% (assuming a true sensitivity of 85%) and a specificity of 75% (assuming a true specificity of 84%), a total sample size of 430 patients is necessary. However, to be able to investigate hypotheses 2 to 4, the sample should include a total of 500 patients, as organ dysfunctions with a low prevalence (lowest with 27% in the case of haematological organ dysfunction) are also investigated.

Diagnosis	**Prevalence** *(absolute)*	**Prevalence** *(relative)*
Systemic Inflammatory Response Syndrome (SIRS)[1]	101/168	60.1%
Hepatic organ dysfunction[1]	51/168	30.4%
Hematologic organ dysfunction[1]	45/168	26.8%
Respiratory organ dysfunction[1]	107/168	63.7%
Renal organ dysfunction[2]	35/168	20.8%
Cardiovascular organ dysfunction[2]	10/168	6.0%

Table 1: Disease prevalence as measured in the CADDIE-2 study sample [2]. These prevalences are used for the disease-specific sample size estimation. All diagnosis with [1] are investigated as primary confirmatory outcomes and those with [2] are investigated as secondary exploratory outcomes.

This study also uses an adaptive caseload planning method [1]. For this purpose, after 50% of the patients (n = 250) have been included, the sample size is recalculated once in a blinded procedure that accounts for the true prevalence measured in this study using the method of Stark and Zapf [1]. If the adaptive case number is greater than the number of patients already included, patients will continue to be included until the adaptive case number is reached (max. 1000 patients; this corresponds to the number of patients in the paediatric intensive care unit within 12 months). If sufficient patients have already been included, recruitment will be terminated. The evaluation itself is based on a non-adjusted error of the first type, because the interim analysis is blinded. Any change in the intended sample size will be communicated.

References

1 Stark M, Zapf A. Sample size calculation and re-estimation based on the prevalence in a single-arm confirmatory diagnostic accuracy study. *Stat Methods Med Res* 2020;**29**. doi:10.1177/0962280220913588

2 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