nature portfolio

Peer Review File

A machine learning and centrifugal microfluidics platform for bedside prediction of sepsis

Corresponding Author: Dr Claudia dos Santos

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

This study presents an innovation by combining machine learning, gene expression analysis, and centrifugal microfluidic technology for the early diagnosis of sepsis. A few comments:

1. It would be helpful to include more detail about the sequence of steps involved in the RNA extraction and ddPCR process. For example, what are the timescales and yields for each step? Are there any specific RNA integrity or efficiency challenges during the automation process?
2. What are the key performance characteristics (e.g., fluidic resistance, temperature stability, material compatibility)? It would be useful to mention how the cartridge's design influences the accuracy and reproducibility of results.
3. Does the observed droplet size distribution consistently result in uniform partitioning and detection across clinical samples?
4. How does fluorescence imaging compare to conventional ddPCR platforms regarding signal-to-noise ratio, detection limits, and sensitivity?
5. What is the level of correlation between the two systems for different types of clinical samples (e.g., ICU patients vs. healthy controls)?
6. Were the models tested on independent cohorts to validate their robustness?
7. How were these HKGs selected, and what makes them suitable for normalisation in sepsis patients? Were there any considerations regarding the stability of these genes under varying physiological conditions (e.g., inflammation, infection)?
8. How consistent is RNA extraction efficiency across different patient populations, and what steps are taken to mitigate variations due to sample handling?
9. How does the Sepset classifier compare to lactate in predicting clinical deterioration within 24 hours? Does the Sepset signature provide any additional diagnostic value over lactate measurements?

Reviewer #2

(Remarks to the Author)
The research article "Using machine learning and centrifugal microfluidics at the point-of-need to predict clinical deterioration of patients with suspected sepsis within the first 24 h" presents a detailed exploration of the Lab-on-Chip device for early sepsis diagnosis within the critical first 24 hours. By leveraging machine learning and the development of an innovative integrated system, this device allows for rapid sepsis detection at the point of care within hours, without requiring specialized personnel. These features render the study particularly relevant to the scientific community. The methodology is well-suited and appropriately designed to meet the study's objectives.
To further enhance the manuscript and emphasize the significance of the developed Lab-on-Chip for early sepsis diagnosis, it would be beneficial to include a discussion in the introduction and discussion/conclusion sections about current sepsis diagnostic methods and the associated timing risks (10.4161/viru.27393; 10.1515/dx-2017-0036).

Reviewer #3

(Remarks to the Author)
I am not in a position to judge the technical details of the methodology's development. As a clinician, however, I find the objective of the study by Malic et al commendable. Early identification of patients with suspected sepsis—especially at true T0, in the emergency department setting—who are at risk of deterioration, is a priority. This is crucial not only for improving disease outcomes but also for optimizing resource utilization in hospitals.
From a clinical perspective, I would like to raise the following points:
1. Cancer, cardiogenic shock, acute coronary syndrome (ACS), and inflammatory bowel disease (IBD) are sepsis mimickers and should be considered in the differential diagnosis of sepsis, particularly in undifferentiated patients (especially in the ED setting). Your cohort of patients with suspected sepsis will likely include individuals with these conditions (15-18% of patients with suspected sepsis are sepsis mimickers and not actually septic). I don't understand why bacterial and viral infections are being used as negative controls. Although these infections may not meet the severity criteria, they are still infections.
2. An AUROC of 0.69 in the ED setting is essentially a negative result, particularly in a setting where prognostication is critically important. A negative predictive value (NPV) of 67% is also a negative result. In the ED, the most important aspect of prognostication is ruling out adverse outcomes.
3. Page 9, line 187: "A SOFA score of 2 was used to discriminate predicted sepsis from non-sepsis." This is unclear. The SOFA score is a prognostic tool, predicting which patients will deteriorate regardless of disease etiology. It is not a diagnostic tool (i.e., distinguishing between sepsis [infection] and sepsis mimickers [non-infectious acute illness]). Predicting the outcome is prognostication, while predicting the etiology is diagnosis.
4. Page 18, line 407: This statement is not supported by the results. Do you truly believe that the implementation of this test could replace the measurement of other clinical variables beyond just suspected sepsis? Many other clinical predictors, which are not measurable through scores (e.g., toxic appearance), could complement this test. Moreover, as mentioned in the limitations, prospective confirmation data are needed. I would be more cautious in considering this test exhaustive for identifying the level of care in undifferentiated patients throughout the manuscript.
5. Page 18, line 417: The concept of misdiagnosis is unclear. Are you referring to the severity of illness in undifferentiated patients, or the etiology of the acute process in undifferentiated patients? Predicting the outcome is prognostication, predicting the etiology is diagnosis.
6. Since the studies by Tom van der Poll, endotyping has been a common research topic. It is clear that, given the turnaround time of the method, a clinician like myself would be very interested in using it. However, I must point out that the sample sizes of the derivation/validation cohorts are not as large as they might seem. I understand that in an observational study, this may not matter much, but it would have been helpful to conduct a power analysis based on the prevalence of outcomes in various settings to identify the appropriate sample sizes for the cohorts.
7. What are the costs of the test per patient? Can you perform a cost analysis?
8. The pathophysiological connection between gene signatures and adverse outcome should perhaps be discussed further.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have addressed most comments. A few additional comments:
1. The manuscript mentions that lactate did not perform well in predicting sepsis in the ER setting. Why this is the case?
2. How do the materials and design impact reproducibility across different clinical samples? (e.g., ICU patients, emergency room patients, healthy controls)
3. can the authors discuss the effect of antibiotics or other treatments on gene expression in sepsis patients?
4. Are there any specific challenges with RNA extraction from patient samples that might affect the performance of the test, particularly in the context of RNA stability?
5. How easily could this system be scaled for use in low-resource settings?

Reviewer #3

(Remarks to the Author)

Point-by-Point Response to Reviewers

Reviewer 1

We thank the reviewer for his time and meaningful comments. We have addressed the comments below and revised the manuscript accordingly.

1. **It would be helpful to include more detail about the sequence of steps involved in the RNA extraction and ddPCR process. For example, what are the timescales and yields for each step? Are there any specific RNA integrity or efficiency challenges during the automation process?**

The detailed protocol for the on-chip RNA extraction and ddPCR was provided in the supplementary information text and accompanying **Supplementary Figure S5A and S5B**. To clarify and direct the reader to the relevant section, we have included in the revised manuscript on lines 693-722 methods section titled: Microfluidic implementation of the SepsetER classifier detection process with references for the **Supplementary Figure S5A and S5B**.

The time for each fluidic step is included in the supplementary tables S8 and S9. The total time for RNA extraction corresponds to 34 min and ddPCR corresponds to 156 min. An extra row in the respective supplementary tables S8 and S9 has been added to better communicate the total time.

To assess the efficiency and reproducibility of the automated RNA extraction protocol, we have compared its performance to the manual extraction process by measuring the total RNA concentration with nanodrop and by performing RT-qPCR on the extracted sample for the six gene targets and two HKGs (n=4). The measured concentrations for the manual and microfluidic methods were similar (6.08±0.38 and 6.87±0.84 ng/uL, respectively), and the Ct values (Figure 1) were also similar for both methods, suggesting similar performance between the two assays in terms of yield and integrity. Standard deviation was slightly lower for the automated protocol, suggesting higher reproducibility between runs.

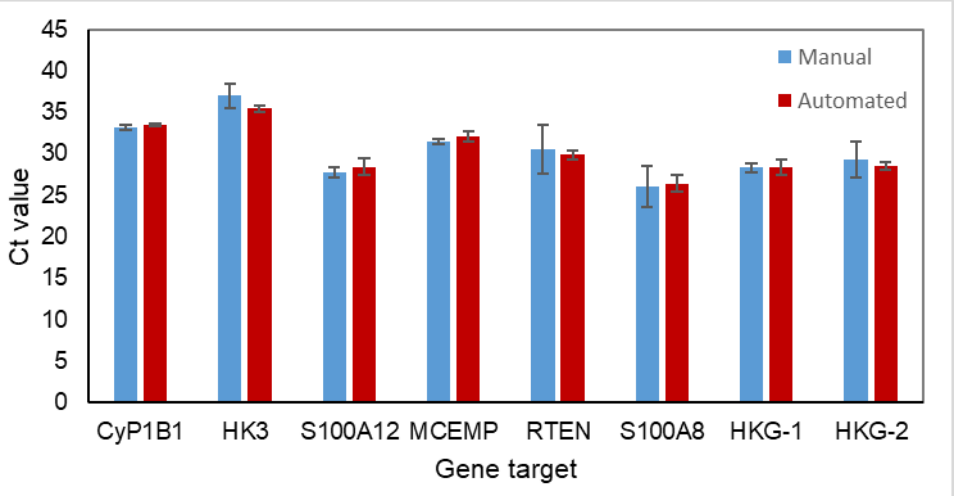


Figure 1. RT-qPCR Ct values of the 6 gene targets and two HKGs obtained for the RNA extracted using automated and manual assays (n=4)

We have added the sentences in the Methods section (line 658):" The extracted RNA was assessed using NanoDrop One spectrophotometer and A260/A280 values were between 1.75 and 2.25, with typical yields in the range of 6–7 ng/µL for both manual and automated protocols." ; as well as line 662:" Ct

values for the 8-gene signature were similar for manual and automated extractions (ΔCt <1) with standard deviation slightly lower for the automated protocol."

2. What are the key performance characteristics (e.g., fluidic resistance, temperature stability, material compatibility)? It would be useful to mention how the cartridge's design influences the accuracy and reproducibility of results.

This is a very important point raised by the reviewer. The two cartridges used for performing sample preparation (RNA extraction) and detection (ddPCR) have very different requirements and stability issues that have been addressed through careful consideration of the features design, materials selection, and fabrication processes. From the material compatibility point of view, both cartridges must not adsorb RNA. The first cartridge for performing RNA extraction is fabricated in Zeonor and the large features required for assay execution (storage reservoirs, RNA extraction chamber and connecting channels) do not impose special fabrication processes (e.g. fabrication can be performed using CNC machining) or assembling procedures (e.g. standard adhesive). On the contrary, the detection cartridge for performing ddPCR requires integration of small microfluidic features and is very sensitivity to delamination or surface tension variabilities. For this specific cartridge, we have found that the combination of PDMS and glass with irreversible bonding capacity performs the best. Please note that the temperature stability is not an issue for either of the devices. For example, the glass transition temperature of Zeonor is ~ 120°C, which is well above the temperature applied during the heating of the lysate and Proteinase K solution (55°C). For the detection cartridge, following droplet generation, the droplets are transferred off chip to a tube contained within a miniature heater module for performing thermal cycling, such that the temperature does not represent an issue.

From the design perspective, the precision of the measurements is increased by carefully tailoring the dimensions of the different features, including droplet generation nozzles and visualization chamber. The size of the nozzles will influence the stability of droplet generation process and resulting monodispersity. The height and the shape of the visualization chamber is important to allow packing of the droplets in a monolayer format for imaging. These parameters have been optimized.

We have revised the manuscript to include on the lines 302-307 a sentence addressing this important point: *"Design of the cartridges and selection of fabrication materials and processes were optimized to ensure compatibility with assay components and reproducible assay execution."*

3. Does the observed droplet size distribution consistently result in uniform partitioning and detection across clinical samples?

This is very important point for accurate sample quantification. The centrifugal step emulsification method employed in this work for droplet generation results in formation of monodisperse droplets across different clinical samples. Please note that following the thermal cycling, some of the droplets will coalesce. However, these droplets will be excluded during the analysis as follows. Image acquisition at the end of the process yields both optical and fluorescence micrographs of droplet monolayers. The analysis software then uses the brightfield image to create a mask that identifies accepted droplets based on their type (dark or light object selection), size and morphology. This mask therefore excludes droplets that have merged during thermal cycling, as well as any artifacts that may be present on the substrate. We have found that amplification efficiency and the resulting fluorescence intensity of positive droplets was similar across all 30 patient samples tested.

4. How does fluorescence imaging compare to conventional ddPCR platforms regarding signal-to-noise ratio, detection limits, and sensitivity?

Fluorescence imaging is more amenable to miniaturization and integration with microfluidic platform, and was the strategy employed to allow the complete assay to be performed using a single instrument. Please also note that fluorescence imaging is used in other ddPCR commercial systems (e.g. Naica from Stilla Technologies). Compared to the laser-induced fluorescence cytometry used in the Bio-Rad system, which likely relies on a photomultiplier tube for detecting fluorescent droplets, fluorescence imaging employs a CMOS camera that may exhibit higher background noise and lower contrast, resulting in a reduced signal-to-noise ratio (SNR). However, for ddPCR application, after 40 cycles, the fluorescence intensity of positive droplets is significantly higher than that for the negative droplets, and can be easily distinguished by fluorescence imaging and image analysis software. The sensitivity and LOD are largely dependent on amplification efficiency, which is in turn affected by the presence of PCR inhibitors in the sample (sample preparation) and the design of the PCR primers and probe. We have optimized both sample preparation as well as primer/probe design in our work, with limit of detection corresponding to 0.1 copy/µL, similar to the Biorad system.

As a direct comparison between Biorad and PREDICT microfluidic platform, the measured values for the SepsetER classifier of the 30 patient samples were input into the prediction algorithm and yielded the same sensitivity and specificity (see Table below).

Platform	N	Sensitivity (%)	Specificity (%)	Precision (%)	NPV (%)	Overall Accuracy (%)
Biorad ddPCR	30	100	89	86	100	**93**
PREDICT ddPCR	30	100	89	86	100	**93**

- Sensitivity – the model correctly predicts the given percent of sepsis patients
- Specificity - the model correctly predicts the given percent of non-sepsis patients
- Precision – out of predicted sepsis samples, the given percent are truly sepsis samples
- NPV – out of predicted sepsis samples, the given percent are truly non-sepsis samples
- Overall Accuracy – percent of predictions that are correct for sepsis and non sepsis samples

5. What is the level of correlation between the two systems for different types of clinical samples (e.g., ICU patients vs. healthy controls)?

We obtained good correlation of the measured expression of the (6) signature genes of interest (GOI) relative to the (2) housekeeping genes for the two systems for different types of clinical samples (please see Supplemental Figure S6). For the 30 patients tested, including ICU patients as well as healthy controls, we obtained identical predictions from the measured output of the SEPSET classifier using both the Biorad ddPCR instrument and the PREDICT microfluidic platform. The absolute counts and normalized values were different between the two platforms, as expected since the two platforms use different initial sample volumes, different sample preparation methods, different levels of multiplexing and different detection methods. The normalized expression (log_2) values were found to be consistently higher for the commercial ddPCR (BioRad) apparatus when compared with the microfluidic platform (CMF). The extent of the difference between the platforms however, was consistent across the different samples ($\leq$2 fold). The differences mitigate with expression level of the respective genes (with differences being lower for the lower expressing GOIs).

Point-by-Point Response to Reviewers

6. Were the models tested on independent cohorts to validate their robustness?

Yes, we tested the Sepset^{ER} genes and algorithm with multiple independent cohorts to validate its robustness (See Table 1). Besides 2nd and 3rd datasets (which are the training/validation datasets), all others (a total of 15 datasets) were independent testing sets: 9 datasets to test sepsis predictions and 6 datasets to test negative control non-sepsis conditions.

7. How were these HKGs selected, and what makes them suitable for normalisation in sepsis patients? Were there any considerations regarding the stability of these genes under varying physiological conditions (e.g., inflammation, infection)?

The selection of HKGs was described in detail in Methods "Sepsis signature and housekeeping gene selection" lines 552-558: *"We analyzed 2,833 housekeeping genes[2] with a discovery cohort of 514 samples to set a baseline for RNA quantity and sequencing depth. We selected housekeeping genes (HKG) with high and consistent expression across all samples based on mean and low variance.* ***We then examined the expression variance of the top 20 HKG candidates across key clinical factors such as age group, gender, sepsis severity, patient location, mortality, etc****. The two housekeeping genes (PTP4A2 and CHTOP) with the lowest variances were used to set a baseline for the SepsetER model."*

8. How consistent is RNA extraction efficiency across different patient populations, and what steps are taken to mitigate variations due to sample handling?

The detailed information regarding on the RNA extraction protocol and efficiency was described previously (https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964(21)00570-3/fulltext). We have included a sentence in the Methods section (line 536) to direct the reader to this reference: "For RNA-Seq, blood collection, RNA extraction and downstream processing was performed as described previously.[32]"*Blood collection and RNA-Seq*

Three millilitres of blood was collected into PAXgene Blood RNA tubes (BD Biosciences) during usual blood collection to ensure intracellular RNA stabilization. RNA isolation, sample preparation, RNA-Seq and downstream processing was performed using well-established standard operating procedures.[17] Total RNA was extracted using PAXgene Blood RNA Kit (Qiagen; Germantown, MD, USA). Quantification and quality measures of total RNA were obtained using an Agilent 2100 Bioanalyzer (Agilent; Santa Clara, USA). Poly-adenylated RNA was captured using NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB; Ipswich, USA). cDNA libraries were prepared using the KAPA Total RNA HyperPrep Kit (Roche; Basel, Switzerland). RNA-Seq was then performed on Illumina Hi-Seq instrument using single read runs of 150 base-pair long sequence reads (excluding adapter/index sequences). RNA-Seq processing and count matrix generation are described in the Supplemental Materials. Of the 363 patients recruited, 348 showed RINs > 65 and had sequencing libraries greater than 1 million (Australia/84, Colombia/67, the Netherlands/104, Vancouver/11, Toronto/82). Severity analysis used all patients, while endotype classification utilized 182 patients in the ER discovery set, 84 patients in the Sydney ER validation cohort and 82 patients in the Toronto ICU validation cohort (Figure S1). Raw sequencing data are available at NCBI GEO (Accession Number GSE185263)."

(More details are available in the Supplementary Methods: https://www.thelancet.com/cms/10.1016/j.ebiom.2021.103776/attachment/d264d12e-9ee8-41c7-9f92-9552a34d1f3a/mmc1.docx)

Overall, our RNA extraction efficiency was excellent and consistent across different patient populations.

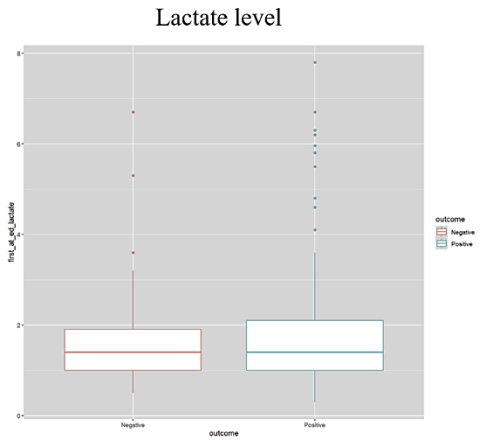
9. How does the Sepset classifier compare to lactate in predicting clinical deterioration within 24 hours? Does the Sepset signature provide any additional diagnostic value over lactate measurements?

Point-by-Point Response to Reviewers

We have lactate data available for 204 out of 488 of our samples (around 42%). The Table below demonstrates a performance analysis with lactate levels vs. SepsetER and the combination (3 models).

Model	Training set	Sepsis (N)	Non-sepsis (N)	Total (N)	Predictors
Sepset^{ER}	Hancock	268	220	488	CR6
Lactate	Hancock_lactate	126	79	205	Lactate level

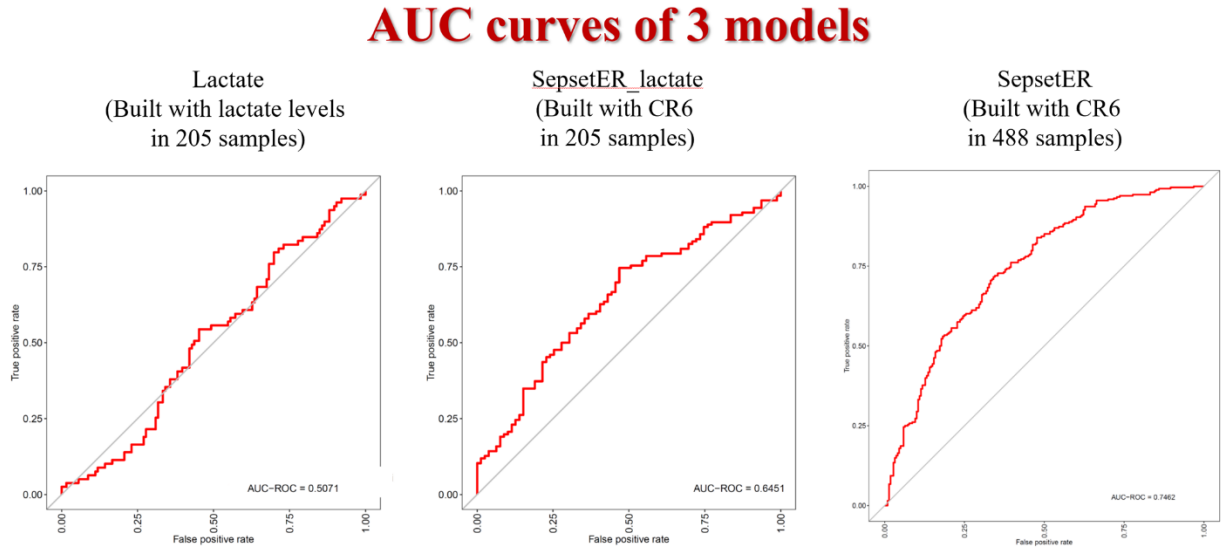
Usually, a model performs better with a larger training set, so for this analysis we retrained the Sepset^{ER} algorithm using the same training set as the lactate model to make the performance comparison fair.



Lactate level showed no significant difference between sepsis (Positive) vs. non-sepsis (Negative) cohorts

Just by looking at the lactate level, it is already possible to tell that model based on lactate will have poor performance since the lactate level lacked discriminating power between sepsis and non-sepsis cohorts. This was confirmed by AUC curves for Hancock ER data which, as expected, showed that the lactate model did not work at all being no better than a random guesses. Sepset^{ER} did very well as shown in our manuscript and the below figure, while combining these two parameters actually decreased the AUC.

Point-by-Point Response to Reviewers



Validated with Hancock emergency department samples:
Lactate model showed AUC = 0.5 which indicate no predictive power at all, aka, random guesses
SepsetER_lactate model showed AUC = 0.65 which suggesting every 2 out of 3 predictions being correct
SepsetER model showed AUC = 0.75 which suggesting every 3 out of 4 predictions being correct

Lactate is usually used as a marker for mortality. It is possible that lactate would have higher prediction value for the development of sepsis later in hospitalization, but in these ER patients it had no discriminatory ability.

We have now included a new Supplemental Figure 7 (Line 1083- 1091) – to describe the performance analysis with lactate levels vs. SepsetER and the combination (3 models).

Reviewer 2

The research article "Using machine learning and centrifugal microfluidics at the point-of-need to predict clinical deterioration of patients with suspected sepsis within the first 24 h" presents a detailed exploration of the Lab-on-Chip device for early sepsis diagnosis within the critical first 24 hours. By leveraging machine learning and the development of an innovative integrated system, this device allows for rapid sepsis detection at the point of care within hours, without requiring specialized personnel. These features render the study particularly relevant to the scientific community. The methodology is well-suited and appropriately designed to meet the study's objectives.

To further enhance the manuscript and emphasize the significance of the developed Lab-on-Chip for early sepsis diagnosis, it would be beneficial to include a discussion in the introduction and discussion/conclusion sections about current sepsis diagnostic methods and the associated timing risks (10.4161/viru.27393; 10.1515/dx-2017-0036).

We thank the reviewer for this comment. We have complied with this request and included the following sentence in the introduction (page 5 Line 96).

"Early diagnosis of sepsis is critical to favourable outcomes[7]. Currently sepsis diagnostic methods include pathogen detection (using conventional or real--time multiplex PCR methods), clinical algorithms based on various tests and risk factors in the patient medical records, and biomarkers including procalcitonin (PCT), C-Reactive Protein (CRP), lactate, white blood cell shape (Intellisep test) as well as a variety of

proposed markers (Calprotectin, MCP-1, Presepsin, etc. Bloos, F., & Reinhart, K. (2013). Rapid diagnosis of sepsis. Virulence, 5(1), 154–160. https://doi.org/10.4161/viru.27393) The proposed international gold standard sequential organ failure assessment (SOFA) score and its early analog quick qSOFA, and more recently a host response molecular test termed Septicyte have also been studied.[refs a and b] Unfortunately, at first clinical presentation, these tests have poor ability to identify patients at risk of developing sepsis."

a) van Oers JAH, de Jong E, Kemperman H, Girbes ARJ, de LangeDW. Diagnostic accuracy 1103 of procalcitonin and c-reactive protein is insufficient to predict proven infection: a 1104 retrospective cohort study in critically ill patients fulfilling the Sepsis-3 criteria. J Appl 1105 Laboratory Med (2020) 5(1):62–72. doi.org/10.1373/jalm.2019.029777

b) Hancock, R.E.W., A. An, C.C. dos Santos and A. HY Lee. 2025. Deciphering sepsis: transforming diagnosis and treatment through systems immunology. Frontiers Science in press (January).

Reviewer #3:

I am not in a position to judge the technical details of the methodology's development. As a clinician, however, I find the objective of the study by Malic et al commendable. Early identification of patients with suspected sepsis—especially at true T0, in the emergency department setting—who are at risk of deterioration, is a priority. This is crucial not only for improving disease outcomes but also for optimizing resource utilization in hospitals.
From a clinical perspective, I would like to raise the following points:

1. Cancer, cardiogenic shock, acute coronary syndrome (ACS), and inflammatory bowel disease (IBD) are sepsis mimickers and should be considered in the differential diagnosis of sepsis, particularly in undifferentiated patients (especially in the ED setting). Your cohort of patients with suspected sepsis will likely include individuals with these conditions (15-18% of patients with suspected sepsis are sepsis mimickers and not actually septic). I don't understand why bacterial and viral infections are being used as negative controls. Although these infections may not meet the severity criteria, they are still infections.

Indeed, we completely agree with the reviewer and felt it was very important to rule out such "sepsis mimickers" as the reviewers suggests. For this reason, Table 1 rows 11-14 includes all of these groups of patients. Based on our analysis of 3171 of these patients we found that the Sepset^{ER} test was not at all able to predict such patients (including more than half who had severe illness), with AUCs ranging from 0.50 to 0.53. We did not per se compare patients with bacterial or viral infections as negative controls. However, many of the patients who were negatively predicted for sepsis (and also many with positive predictions for sepsis) had bacterial or viral infections and our accurate predictions in these patients was borne out by the correlation of a positive Sepset^{ER} prediction and SOFA score $\geq$2. For example, all patients in the COVID-19 sepsis cohort were admitted to hospital with viral (COVID-19) infections but only those with SOFA score $\geq$2 gave accurate predictions (AUC = 0.85) for sepsis.

2. An AUROC of 0.69 in the ED setting is essentially a negative result, particularly in a setting where prognostication is critically important. A negative predictive value (NPV) of 67% is also a negative result. In the ED, the most important aspect of prognostication is ruling out adverse outcomes.

We thank the reviewer for this point. We understand what the reviewer is saying but it is worth putting our prediction in the context of all other test to date that have sensitivity and specificity of 50% (as described in the manuscript). Nevertheless specificity (correct prediction of sepsis) is 70%, i.e. 7 of 10 patients are correctly predicted. This accuracy increases by the time a patient hits the ICU to around 9 of 10 patients correctly predicted. We would anticipate that patients with a positive test or developing

symptoms should be retested at a later time (around 6-12 hours of hospitalization based on our preliminary analyses), with the likelihood of greater accuracy. The limited accuracy early on is likely due to the fact that some patients undergo early interventions (e.g. early antibiotics or anti-virals) that prevent sepsis from developing, as stated in the manuscript (lines 411-413). We agree that our specificity/NPV are lower in ED patients so the test early on is not effective in ruling out sepsis. But these numbers also improve substantially in later stage (ICU) patients. We have included a sentence in the discussion (line 413) to discuss this point:" This accuracy increases upon ICU admission suggesting that patients with a positive test or developing symptoms should be retested at a later time (around 6-12 hours of hospitalization based on our preliminary analyses), with the likelihood of greater accuracy. "

3. Page 9, line 187: "A SOFA score of 2 was used to discriminate predicted sepsis from non-sepsis." This is unclear. The SOFA score is a prognostic tool, predicting which patients will deteriorate regardless of disease etiology. It is not a diagnostic tool (i.e., distinguishing between sepsis [infection] and sepsis mimickers [non-infectious acute illness]). Predicting the outcome is prognostication, while predicting the etiology is diagnosis.

Sorry we meant to state "A SOFA score of 2 was used to discriminate sepsis from non-sepsis." The sentence was changed accordingly. This is not a diagnostic statement but rather is the international gold standard (Sepsis 3) definition. Prior to Sepsis 3, sepsis was split into sepsis and severe sepsis with the latter patients showing organ failure/deterioration. Now severe sepsis is redefined as sepsis. This SOFA is the outcome we measure and it allows us to distinguish between sepsis (which please note is not the same as infection) and all other conditions that are not sepsis. We respectfully submit, the primary purpose of the SOFA score is to objectively describe organ (dys)function rather than to predict outcome, so no associated equation developed for prediction is currently in use. (Moreno, R. et al. The Sequential Organ Failure Assessment (SOFA) Score: has the time come for an update? Critical Care 27, 15 , 2023). The score was designed to be easy to use and had to fulfill several guiding principles: 'It should be a descriptor of individual organ dysfunction during the ICU stay and not a predictor of vital status at hospital discharge for heterogeneous groups of patients…' (Crit Care Sci. 2024 Mar 25;36:e20240296en. doi: 10.62675/2965-2774.20240296-en).

4. Page 18, line 407: This statement is not supported by the results. Do you truly believe that the implementation of this test could replace the measurement of other clinical variables beyond just suspected sepsis? Many other clinical predictors, which are not measurable through scores (e.g., toxic appearance), could complement this test. Moreover, as mentioned in the limitations, prospective confirmation data are needed. I would be more cautious in considering this test exhaustive for identifying the level of care in undifferentiated patients throughout the manuscript.

We have amended this statement, adding the underlined words as suggested by the reviewer, to read: "This means that, other than suspicion of sepsis, no other information should be needed for accurate prediction of the clinical course over the first 24 h following initial assessment, although other clinical predictors could complement this test."

5. Page 18, line 417: The concept of misdiagnosis is unclear. Are you referring to the severity of illness in undifferentiated patients, or the etiology of the acute process in undifferentiated patients? Predicting the outcome is prognostication, predicting the etiology is diagnosis.

The statement should have read accuracy instead of specificity: i.e. "Poor accuracy of current syndromic diagnostic approaches results in unnecessary use of drugs (e.g., antibiotics and steroids) and exposure to harm (iatrogenicity) associated with hospital and ICU-level care." We have corrected the statement accordingly. Accuracy is used as percent of patients with a correct diagnosis (sepsis or non-sepsis). As described in the text, current tests are only 50% accurate, especially early on. Thus, the physician would

be encouraged to use potent antibiotics and steroids for patients who do not have sepsis, while withholding these for some patients who do have sepsis.

6. Since the studies by Tom van der Poll, endotyping has been a common research topic. It is clear that, given the turnaround time of the method, a clinician like myself would be very interested in using it. However, I must point out that the sample sizes of the derivation/validation cohorts are not as large as they might seem. I understand that in an observational study, this may not matter much, but it would have been helpful to conduct a power analysis based on the prevalence of outcomes in various settings to identify the appropriate sample sizes for the cohorts.

Thanks for this. It is important to note that the actual Sepset^{ER} test was established using conventional PCR as described in Table 1 which shows that the cohort sizes were very large. We did a power analysis to start this study. We found that, using DeLong's method on the area under the curve of a receiver operating characteristic curve, assuming a power of 90% and using a two-sided alpha of 5%, one would need 376 subjects to demonstrate superior balanced accuracy to a common test like PCT. Our initial test was in fact established and validated using 586 patients. Once the test was established, we demonstrated that the ddPCR method gave similar results as revealed in the bottom 3 rows of Table 1.

7. What are the costs of the test per patient? Can you perform a cost analysis?

At this stage performing a detailed cost analysis is moot since costs decrease when a test is mass produced. We have estimated an initial cost to the manufacturer of conventional RNA extraction and PCR of ~$16 per test; operator costs would more than double this.

For the PREDICT assay, we estimate the total cost of the consumables (cartridges and reagents) to be $20 for pilot production (see Table 1). The cost of the prototyping PREDICT microfluidic platform instrument is $40,000, which is expected to decrease in the production phase.

Table 1. Estimated cost of the PREDICT microfluidic platform instrument and single test consumables

Component	Cost ($)
Prototype instrument	40,000
Consumables	
Cartridges	6
Reagents	
RNA extraction (magnetic)	6
ddPCR (oil and reagents)	8
Total Platform	40,000
Total consumables (per test)	19

However, please note that these numbers are early estimates and perhaps outside of the scope of our article.

8. The pathophysiological connection between gene signatures and adverse outcome should perhaps be discussed further.

We thank the reviewer for this point. We have added a sentence in the discussion (line 450) "Since Sepset^{ER} is based on a gene expression analysis of immunosuppression/ cellular reprogramming, there is likely a very strong relationship between a positive (for sepsis) assay and the pathophysiological

consequences of a dysfunctional host response to expression as reinforced by the relationship of a positive Sepset^{ER} test and organ dysfunction (SOFA score).”

Thank you, Reviewer 1, for your comments. Please find below our point-by-point responses and details regarding modifications made to the manuscript based on excellent suggestions for clarification.

Responses to reviewer 1

1. The manuscript mentions that lactate did not perform well in predicting sepsis in the ER setting. Why this is the case?

Our Response: We thank the reviewer for raising this question. Elevated serum lactate measures acidosis and is used as a general test for serious health problems. While often used for sepsis, it is most associated with an increase in 24-hr and 30-day mortality in Emergency Department (ED) patients with severe sepsis and septic shock (PMID: 32807232). Thus, it is an indicator of mortality not severe sepsis and organ failure per se, whereas our test predicts the onset of sepsis and organ failure (SOFA score ≥ 2). Overall, only 23% of sepsis patients die, which is why prediction of severe sepsis and mortality are in fact separate. Indeed, in our eBiomedicines paper[33], mortality and Sepset signatures were quite distinct. Thus, we have added after the mention of lactate: "Lactate is used as a general test for serious health problems, and while used in sepsis, tends to indicate mortality risk[60], which is separate from progression to sepsis[33]." (page 17 lines 305-6)

2. How do the materials and design impact reproducibility across different clinical samples? (e.g., ICU patients, emergency room patients, healthy controls).

Our Response: We wrote in Discussion (lines 433-6): "We have clearly shown that both conventional RT-qPCR and ddPCR, as well as the PREDICT platform, give very similar accuracy results in all of these three groups. So, we can conclude that the test is quite robust and materials and design have at most a modest impact."

3. Can the authors discuss the effect of antibiotics or other treatments on gene expression in sepsis patients?

Our Response: We have modified to add more detail to the previous sentence (page 10 lines 350-4): "In our previous study[33] we demonstrated that despite the use of antibiotics in most patients, the Sepset signature was robust in determining progression to sepsis. However, in ER patients the sensitivity was lower (70%) than in early ICU patients. It is possible this reflected to some extent that in some ER patients, the early use of antibiotics actually prevented progression to sepsis and organ failure as defined by SOFA score ≥ 2."

4. Are there any specific challenges with RNA extraction from patient samples that might affect the performance of the test, particularly in the context of RNA stability?

Our Response: The blood collection was performed in RNA-stabilized collection system (PAXgene blood RNA tubes). We have tried many commercial extraction methods, both magnetic-based and column based (both, in the context of conventional RT-qPCR, as well as during PREDICT system implementation) some of which performed better than others (in terms of yield and purity). Since all of these use buffers that prevent RNA degradation (i.e. maintain RNA stability), we have concluded that the key discriminator is the purity of the RNA which can impact on the CT curves that are used to measure relative RNA concentration.

5. How easily could this system be scaled for use in low-resource settings?

Our Response: At this time, resources and cost of equipment are a major issue for low resource settings and these issues will have to be overcome for low-resource settings. Our lab is currently building a system based on our centrifugal microfluidics platform for deployment in remote communities (e.g. Canadian North – Manitoba, for food safety) and has been qualified for remote deployment on the International Space Station (ISS) in 2025 for nucleic acid and protein extraction built by MDA Robotics. In addition, we are working on the development of a second generation PREDICT system (both platform as well as cartridges for full automation), this modified system could be deployed in low resource settings.