A multi-platform approach to identify a blood-based host protein signature for distinguishing between bacterial and viral infections in febrile children (PERFORM): a multi-cohort machine learning study

Heather R Jackson, Judith Zandstra*, Stephanie Menikou, Melissa Shea Hamilton, Andrew J McArdle, Roman Fischer, Adam M Thorne, Honglei Huang, Michael W Tanck, Machiel H Jansen, Tisham De, Philipp K A Agyeman, Ulrich Von Both, Enitan D Carrol, Marieke Emonts, Irini Eleftheriou, Michiel Van der Flier, Colin Fink, Jolein Gloerich, Ronald De Groot, Henriette A Moll, Marko Pokorn, Andrew J Pollard, Luregn J Schlapbach, Maria N Tsolia, Effua Usuf, Victoria J Wright, Shunmay Yeung, Dace Zavadska, Werner Zenz, Lachlan J M Coin, Climent Casals-Pascual, Aubrey J Cunnington, Federico Martinon-Torres, Jethro A Herberg, Marien I de Jonge†, Michael Levin†, Taco W Kuijpers†, Myrsini Kaforou†, on behalf of the PERFORM consortium‡*

Summary

Background Differentiating between self-resolving viral infections and bacterial infections in children who are febrile is a common challenge, causing difficulties in identifying which individuals require antibiotics. Studying the host response to infection can provide useful insights and can lead to the identification of biomarkers of infection with diagnostic potential. This study aimed to identify host protein biomarkers for future development into an accurate, rapid point-of-care test that can distinguish between bacterial and viral infections, by recruiting children presenting to health-care settings with fever or a history of fever in the previous 72 h.

Methods In this multi-cohort machine learning study, patient data were taken from EUCLIDS, the Swiss Pediatric Sepsis study, the GENDRES study, and the PERFORM study, which were all based in Europe. We generated three high-dimensional proteomic datasets (SomaScan and two via liquid chromatography tandem mass spectrometry, referred to as MS-A and MS-B) using targeted and untargeted platforms (SomaScan and liquid chromatography mass spectrometry). Protein biomarkers were then shortlisted using differential abundance analysis, feature selection using forward selection-partial least squares (FS-PLS; 100 iterations), along with a literature search. Identified proteins were tested with Luminex and ELISA and iterative FS-PLS was done again (25 iterations) on the Luminex results alone, and the Luminex and ELISA results together. A sparse protein signature for distinguishing between bacterial and viral infections was identified from the selected proteins. The performance of this signature was finally tested using Luminex assays and by calculating disease risk scores.

Findings 376 children provided serum or plasma samples for use in the discovery of protein biomarkers. 79 serum samples were collected for the generation of the SomaScan dataset, 147 plasma samples for the MS-A dataset, and 150 plasma samples for the MS-B dataset. Differential abundance analysis, and the first round of feature selection using FS-PLS identified 35 protein biomarker candidates, of which 13 had commercial ELISA or Luminex tests available. 16 proteins with ELISA or Luminex tests available were identified by literature review. Further evaluation via Luminex and ELISA and the second round of feature selection using FS-PLS revealed a six-protein signature: three of the included proteins are elevated in bacterial infections (SELE, NGAL, and IFN-γ), and three are elevated in viral infections (IL18, NCAM1, and LG3BP). Performance testing of the signature using Luminex assays revealed area under the receiver operating characteristic curve values between 89·4% and 93·6%.

Interpretation This study has led to the identification of a protein signature that could be ultimately developed into a blood-based point-of-care diagnostic test for rapidly diagnosing bacterial and viral infections in febrile children. Such a test has the potential to greatly improve care of children who are febrile, ensuring that the correct individuals receive antibiotics.

Funding European Union's Horizon 2020 research and innovation programme, the European Union's Seventh Framework Programme (EUCLIDS), Imperial Biomedical Research Centre of the National Institute for Health Research, the Wellcome Trust and Medical Research Foundation, Instituto de Salud Carlos III, Consorcio Centro de Investigación Biomédica en Red de Enfermedades Respiratorias, Grupos de Refeencia Competitiva, Swiss State Secretariat for Education, Research and Innovation.

Lancet Digit Health 2023; 5: e774–85

*Joint first authors

†Joint last authors

‡Study group members are listed in the appendix

Section of Paediatric Infectious Disease, Faculty of Medicine, and Centre for Paediatrics and Child Health, Imperial College London, London, UK (H R Jackson PhD, S Menikou PhD, M S Hamilton PhD, A J McArdle MB Bchir, T De PhD, V J Wright PhD, Prof A J Cunnington PhD, J A Herberg PhD, Prof M Levin PhD, M Kaforou PhD); **Sanquin Research and Landsteiner Laboratory, Department of Immunopathology, Sanquin Blood Supply** (J Zandstra MSc, Prof T W Kuijpers PhD), **Department of Pediatric Immunology, Rheumatology and Infectious Diseases, Emma Children's Hospital** (J Zandstra, Prof T W Kuijpers, M H Jansen BSc), **and Department of Epidemiology and Data Science** (M W Tanck PhD), **Amsterdam University Medical Center (UMC), Amsterdam, Netherlands; Discovery Proteomics Facility** (R Fischer PhD), **Target Discovery Institute** (H Huang PhD), **Nuffield Department of Medicine, University of Oxford, Oxford, UK; Department of Surgery, Section of Hepatobiliary Surgery and Liver Transplantation, University of Groningen, University Medical Center Groningen, Groningen, Netherlands** (A M Thorne BSc);

Department of Pediatrics, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland (P K A Agyeman MD); Infectious Diseases, Department of Pediatrics, Dr von Hauner Children's Hospital, University Hospital, LMU Munich, Munich, Germany (U Von Both MD); Department of Clinical Infection Microbiology and Immunology, University of Liverpool Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Liverpool, UK (Prof E D Carrol MD); Paediatric Infectious Diseases and Immunology Department, Newcastle upon Tyne Hospitals Foundation Trust, Great North Children's Hospital, Newcastle upon Tyne, UK (Prof M Emonts PhD); Second Department of Paediatrics, National and Kapodistrian University of Athens (NKUA), School of Medicine, Panagiotis & Aglaia, Kyriakou Children's Hospital, Athens, Greece (I Eleftheriou MD, Prof M N Tsolia PhD); Paediatric Infectious Diseases and Immunology, Wilhelmina Children's Hospital, University Medical Center Utrecht, Utrecht, Netherlands (M Van der Flier PhD); Pediatric Infectious Diseases and Immunology Amalia Children's Hospital (M Van der Flier), Laboratory of Infectious Diseases (M Van der Flier), Translational Metabolic Laboratory (J Gloerich PhD, Prof R Se Groot PhD, Prof M I de Jonge PhD), and Laboratory of Medical Immunology (Prof M I de Jonge), Department of Laboratory Medicine, Radboud Institute of Molecular Life Sciences, Radboud UMC, Nijmegen, Netherlands; Micropathology, University of Warwick, Warwick, UK (Prof C Fink PhD); Department of Pediatrics, Erasmus MC, Rotterdam, Netherlands (Prof H A Moll MD); Division of Paediatrics, University Medical Centre Ljubljana and Medical Faculty, University of Ljubljana, Ljubljana, Slovenia (M Pokorn PhD); Oxford Vaccine Group Department of Paediatrics, University of Oxford and the NIHR Oxford Biomedical Research Centre,

Research in context

Evidence before this study

The majority of children who are febrile attending health-care settings have self-resolving viral infections; however, a small minority have bacterial infections that can be life-threatening. Clinical features alone do not reliably distinguish between bacterial and viral infections. Culture-based methods are the best diagnostic approaches; however, they have various shortcomings including slow turnaround times, low sensitivity, and high resource intensity. We searched PubMed for papers published between database inception and June 6, 2023 using the search terms "bacterial" AND "viral" AND ("paediatric" OR "pediatric" OR "children") AND ("proteomics" OR "host protein" OR "host protein abundance") AND ("signature" OR "diagnosis"). Our search returned 24 papers, with 11 papers reporting on the performance of a single three-protein signature (C-reactive protein, interferon γ-induced protein 10 [IP10] and tumour necrosis factor (TNF)-related apoptosis-inducing ligand [TRAIL]) for distinguishing between bacterial and viral infections, using various study populations to evaluate its performance. Although the three-protein signature has promising performance in paediatric populations, the proteins comprising the signature were not identified from child protein profile data and were instead identified from literature searches or adult transcriptomic data. A further two studies were identified, which explored the performance of distinct protein signatures using proteins selected from the literature (signature one: C-reactive protein, procalcitonin, IL-6, NGAL, MxA, TRAIL, and IP-10 and signature two: procalcitonin, TRAIL, IL-4, IL-6, CXCL10, IFN-γ, and LCN2) in distinguishing bacterial infections from viral infections in emergency departments in France and Spain. As such, there might be further optimal host proteins for distinguishing between bacterial and viral infections in children who are febrile.

Added value of this study

This study is the first to use high-dimensional proteomic datasets composed of children from a range of countries and health-care settings to identify a host protein signature for distinguishing between bacterial and viral infections. We have identified a combination of six proteins (six-protein signature) that has not yet been reported as a combination of proteins with diagnostic potential for distinguishing between bacterial and viral infections in children who are febrile. Furthermore, we have identified a list of novel host protein biomarkers for differentiating between bacterial and viral infections in children who are febrile that have not yet been reported in the literature. This study also contributes three high-quality, high-dimensional proteomic datasets that are publicly available for reuse.

Implications of all the available evidence

Host protein biomarkers have clear potential to improve the diagnosis of bacterial and viral infections in children who are febrile. The six-protein signature presented here could be developed into a rapid blood-based point-of-care diagnostic test with high accuracy, considerably improving the care of children who are febrile by reducing unnecessary antibiotic administration. Using an easy-to-access fluid such as blood would make obtaining samples simpler compared with the current standard for infectious disease diagnosis, in which samples are sourced from the site of infection, thus increasing the applicability and ease of use of such a diagnostic test.

Introduction

Most children who are febrile and attending health-care settings have self-resolving viral infections; however, a small minority have bacterial infections, which can be life-threatening if left untreated. Clinical features do not reliably distinguish between bacterial and viral infections,[1] with confirmed bacterial infections currently identified through culture tests from normally sterile sites. The results can take several days to become available and can be unreliable so antibiotics are often given on an empirical basis, contributing to the spread of antimicrobial resistance.[2] Conversely, severe bacterial infections can be missed, which can have life-threatening consequences. In addition to culture-based diagnostic tests, blood biomarkers such as C-reactive protein and procalcitonin are often used as markers of bacterial infection.[3,4] Despite their frequent use, C-reactive protein and procalcitonin are imperfect biomarkers for distinguishing bacterial from viral infections, as elevated concentrations of both biomarkers have not only been observed in the plasma of patients with confirmed bacterial infections but also in those with viral infections, including SARS-CoV-2, and many non-infectious conditions.[5]

A rapid, accurate, point-of-care test is urgently required for distinguishing between bacterial and viral infections in children who are febrile. Interrogation of host proteomic profiles obtained from individuals with infectious and inflammatory diseases offers unique insights into disease pathogenesis and can reveal novel protein biomarkers with diagnostic potential.[6] Multiple host protein biomarker candidates for diagnosing febrile illness in children have been identified,[7,8] with one protein signature (ie, combination of proteins) proposed and developed into a commercialised point-of-care diagnostic test: MeMed BV (MeMed, Tirat Carmel).[9] MeMed BV uses three host protein biomarkers: C-reactive protein, interferon γ-induced protein 10 (IP10), and tumour necrosis factor (TNF)-related apoptosis-inducing ligand (TRAIL). The three-protein signature included in MeMed BV was identified through hypothesis-driven literature searches and targeted screening of biomarker candidates,[9] and developed and optimised for use in populations of all ages. Although

the test has promising performance in paediatric populations,[10] there might be alternative protein biomarkers that are superior for diagnosing bacterial and viral infections in children who are febrile.

We aimed to identify protein biomarkers for diagnosing febrile illness in children and, in combination with literature-derived proteins, identify and evaluate the best reduced combinations of protein biomarkers for distinguishing between bacterial and viral infections.

Methods

Study design

In this multi-cohort machine learning study, we used samples from patients recruited prospectively into the EUCLIDS study, the Swiss Pediatric Sepsis study, the GENDRES study, and the PERFORM study (appendix p 1).[11,12] These studies recruited children from across 16 European hospitals and used a previously validated phenotyping algorithm and sampling procedures.[11,12] The studies enrolled diverse yet well phenotyped children with suspected infection. Recruitment dates ranged from Sept 1, 2011, to Dec 31, 2019.

Participants

All participants were aged 18 years and younger and were required to have a fever or history of fever (temperature ≥38·0°C) in the previous 72 h for inclusion. Patients defined as having definite viral infections were required to have an identified virus that matched the clinical syndrome, in addition to C-reactive protein concentrations of 60 mg/L or less, and reduced numbers of neutrophils at 12×10^{9}/mL or less. Patients defined as having definite bacterial infections were required to have a bacterial pathogen identified from a sterile site that matched the clinical syndrome (appendix pp 1, 11).

Parental informed consent and assent from older children was collected at the time of recruitment into each respective dataset. Ethical approval was obtained at the coordinating site (Imperial College London, 16/LO/1684) and separately at each participating centre.

The healthy control group from PERFORM constituted children younger than 18 years who were not febrile and who had no symptoms of infection. These children had blood tests for reasons unrelated to infection or inflammation and had not received vaccinations within the preceding 3 weeks.

Procedures

Serum and plasma samples were obtained from participants in the EUCLIDS study, the Swiss Pediatric Sepsis study, and the GENDRES study, and plasma samples were obtained from participants in the PERFORM study. Samples were taken as part of the recruitment process by a clinical team member and were transported and stored at –80°C. Serum samples used in the SomaScan experiment were handled by SomaLogic (Boulder, CO, USA), plasma samples used in the liquid chromatography tandem mass spectrometry experiments were handled by the Discovery Proteomics Facility (Oxford, UK).

We did a robust protein biomarker identification study (ie, the high-throughput screening phase). Following the high-throughput screening phase, the most significant or most frequently selected biomarker candidates were quantified using Luminex and ELISA, which are similar to the type of platform used in a point-of-care diagnostic test (ie, the signature refinement phase). We then validated the final protein biomarker signature (ie, the signature validation phase).

The high-throughput screening phase involved identifying potential protein biomarker candidates for distinguishing between samples obtained from patients with definite bacterial and definite viral infections. Three separate datasets were generated for the discovery of protein biomarkers. The SomaScan dataset was generated from serum samples using the multiplexed SomaScan aptamer-based platform (SomaLogic; 1.3K Assay). The remaining two datasets were generated from plasma samples using liquid chromatography tandem mass spectrometry, to be referred to as the MS-A and MS-B dataset (appendix p 2). The MS-A dataset used plasma samples from patients recruited into the EUCLIDS study, and the MS-B dataset used plasma samples from patients recruited into the PERFORM study. Aside from cohort, the data generation protocols were identical for MS-A and MS-B. Differential abundance analysis was done on each dataset to identify lists of proteins that were significantly differentially abundant between patients with definite bacterial and definite viral infections. Next, feature selection was done on these datasets to identify small protein signatures, (ie, combinations of proteins with diagnostic potential; appendix p 3). For machine learning, an in-house feature selection method, forward selection-partial least squares (FS-PLS),[1,13] was applied to each high-throughput dataset to identify protein signatures for differentiating between definite bacterial and definite viral infections (appendix p 3). FS-PLS was applied across 100 iterations to each dataset, each time with a different training and test split at a ratio of 7:3. This approach was used to enable identification of the most robust proteins in addition to the best diagnostic combination of proteins.

We also did a literature search of PubMed on Dec 1, 2017 (appendix p 3), to explore studies published from Jan 1, 2005, that reported biomarkers for diagnosing bacterial and viral infections. The search used the terms "infection, bacterial or viral, biomarker, plasma or serum" and additional search terms including "biomarker", "cytokine", "chemokine", "growth factor", and "multiplex" or "Luminex". The goal of the literature search was to identify known protein biomarkers for evaluation of their performance in our cohort of patients.

Protein biomarker candidates identified in the high-throughput screening phase were considered for the

Oxford, UK (Prof A J Pollard FmedSci); **Department of Intensive Care and Neonatology and Children's Research Center, University Children's Hospital Zurich, Zurich, Switzerland** (L J Schlapbach PhD); **Child Health Research Centre, The University of Queensland, Brisbane, NSW, Australia** (L J Schlapbach); **Medical Research Council Unit The Gambia at the London School of Hygiene & Tropical Medicine, Fajara, Gambia** (E Usuf PhD); **Clinical Research Department, Faculty of Infectious and Tropical Disease, London School of Hygiene & Tropical Medicine, London, UK** (Prof S Yeung PhD); **Children's Clinical University Hospital, Rīga Stradins University, Rīga, Latvia** (D Zavadska PhD); **University Clinic of Paediatrics and Adolescent Medicine, Department of General Paediatrics, Medical University Graz, Graz, Austria** (W Zenz MD); **Department of Microbiology and Immunology, University of Melbourne at The Peter Doherty Institute for Infection and Immunity, Melbourne, VIC, Australia** (Prof L J M Coin PhD); **Department of Clinical Microbiology, CDB, Hospital Clínic of Barcelona, University of Barcelona, Barcelona, Spain** (C Casals-Pascual PhD); **Translational Pediatrics and Infectious Diseases Section, Pediatrics Department** (Prof F Martinon-Torres PhD) **and Genetics, Vaccines, Infectious Diseases, and Pediatrics research group GENVIP, Instituto de Investigación Sanitaria de Santiago (IDIS)** (F Martinon-Torres), **Universidade de Santiago de Compostela (USC), Santiago de Compostela, Spain; Consorcio Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (CIBERES), Instituto de Salud Carlos III, Madrid, Spain** (F Martinon-Torres)

Correspondence to:
Dr Myrsini Kaforou, Section of Paediatric Infectious Disease, Faculty of Medicine, and Centre for Paediatrics and Child Health, Imperial College London, London W2 1NY, UK
m.kaforou@imperial.ac.uk

See **Online** for appendix

signature refinement phase in addition to the proteins identified from the literature. Multiple inclusion criteria were used to introduce redundancy in case some proteins did not successfully translate across platforms (appendix p 3).

In the signature refinement phase, concentrations of protein were measured by ELISA or Luminex immunoassays in an independent set of plasma samples from the PERFORM study of children with definite bacterial and definite viral infection, using standard curves to identify protein concentrations (appendix p 4). Iterative FS-PLS was applied to the proteins measured using the Luminex results alone (Luminex signature), and then to the proteins measured using both the ELISA and Luminex results (Luminex and ELISA signature). The aim of the signature refinement phase was to narrow down the list of protein biomarkers identified in the high-throughput screening phase and from the literature screening phase, to identify the optimal combination of proteins that, when measured using a targeted, simpler platform, have the best performance at distinguishing between bacterial and viral infections. In this phase, a small protein signature for distinguishing between bacterial and viral infections was identified from the selected proteins.

The performance of this signature was tested in a further independent cohort of patients recruited into the PERFORM study in the signature validation phase with protein concentrations measured using Luminex immunoassays (appendix pp 4–5). The protein signatures identified in the signature refinement phase as the optimal predictive signature for distinguishing between bacterial and viral infections were taken forward to the signature validation phase. Proteins were measured on plasma samples from PERFORM patients with definite bacterial and definite viral infections, non-sterile definite bacterial infections, probable bacterial infections, bacterial syndrome, probable viral infections, viral syndrome, and the healthy control group. The performance of the signature identified in the signature refinement phase in differentiating between definite bacterial and definite viral infection was evaluated in all phenotypic groups and the performance of the signature was compared with the performance of the combination of proteins included in MeMed BV (composed of C-reactive protein, IP10, and TRAIL; appendix p 5), but was not measured on the MeMed BV platform.[9] The healthy control group data were used purely for data pre-processing purposes.

The performance of the signature identified in the signature validation phase was evaluated in terms of its ability to classify patients with definite bacterial infections and definite viral infections. A disease risk score was generated and the optimal threshold above or below which a patient was classified as definite bacterial was obtained, which corresponded to the maximal specificity with sensitivity greater than 90%, to prioritise correct classification of definite bacterial infections. There was no indeterminate classification. The signature was also tested on the patients in the following groups: non-sterile definite bacterial infection, probable bacterial infection, bacterial syndrome, probable viral infection, and viral syndrome. However, these groups were not used for the calculation of performance metrics. A disease risk score was generated for these patients and they were classified as either definite bacterial or definite viral using the threshold calculated previously. Retrospective clinical classifications (eg, definite bacterial) was available for all participants throughout this process for evaluation of test performance.

At each phase of the study, independent cohorts of patients were used, with no overlap between any patients included in datasets generated at any stage of the study. Biological sex at birth as recorded in patients' hospital registration information was used.

Statistical analysis

All statistical analyses in this study were done using R (version 3.6.1).[14] Normalisation (appendix pp 2–3) and analytical processes were done on the three high-throughput discovery proteomic datasets independently due to differences in sample type, study cohort composition, and quantification platform. Limma[15] was used for differential abundance analysis to identify proteins significantly differentially abundant between children with definite bacterial and definite viral infections. Age and sex were included as covariates for all three datasets, with plate as an additional covariate for the SomaScan dataset. P values were adjusted using the Benjamini-Hochberg procedure,[16] with a significance level of 0·05.

In the high-throughput screening phase, FS-PLS was done iteratively on 100 different training:test splits (ratio of 70:30) of each dataset (appendix p 3). For each iteration, the signature identified by FS-PLS with the highest area under the receiver operating characteristic curve (AUC ROC) in the test dataset was taken forward. The frequency with which each signature and each individual protein were selected across the 100 iterations was calculated. A robustness value was calculated as the number of times a protein was selected across all iterations, divided by the total number of iterations. Following the high-throughput screening phase, a shortlist of potential protein biomarkers for distinguishing between bacterial and viral infections was identified.

In the signature refinement phase, levels of proteins were first compared between definite bacterial and definite viral patients using the Mann-Whitney U test. FS-PLS was then run twice, either on proteins measured using ELISA and Luminex and then just on the proteins measured using Luminex. All parameters were the same as the parameters used in the high-throughput screening phase, except the number of iterations which was reduced

to 25 to reflect the lower number of dimensions. For all samples, a weighted disease risk score[1] was calculated for both signatures by multiplying the abundance values of each protein by their weights (also known as coefficients) from FS-PLS (appendix p 4). From the disease risk score, the AUCs and partial AUCs were calculated[17] for the two signatures at 90% sensitivity and specificity, 95% sensitivity and specificity, and maximal sensitivity and specificity using Youden's index[18] were also calculated to contrast the performance of the two signatures. Comparisons were to identify which was the optimal signature. To identify proteins that could distinguish between bacterial and viral infections when C-reactive protein is low in bacterial infections, differential abundance analysis was done using Limma between definite bacterial samples with C-reactive protein of 60 mg/L or less and definite viral samples. Age and sex were included as covariates in the model.

In the signature validation phase, weighted disease risk scores were calculated using the FS-PLS model weights from the signature refinement phase (original weights) along with retrained model weights from the general linear models using the signature validation phase data (appendix pp 4–5). A simple disease risk score was also used (appendix p 4),[1] which entails adding the total abundance of the over-expressed proteins and subtracting the total abundance of the under-expressed proteins. The signature performance was again evaluated through calculating AUCs and partial AUCs at 90% sensitivity and specificity, 95% sensitivity and specificity, and the maximal sensitivity and specificity were calculated using Youden's index.[18] Disease risk scores were also calculated for patients in other phenotypic groups (ie, not definite bacterial or definite viral) and individuals and samples were classified using the disease risk score threshold corresponding to the maximal specificity with a sensitivity greater than 90%. Patients were classified in a binary manner (ie, definite bacterial or definite viral) with no indeterminate classification.

Role of the funding source

The funders of the study had no role in the study design, data collection, data analysis, data interpretation, or writing of the report.

Results

376 children provided serum or plasma samples for the identification of protein biomarkers for bacterial infection, with 79 providing serum for the generation of the SomaScan dataset, 147 providing plasma for the generation of the MS-A dataset, and 150 providing plasma for the generation of the MS-B dataset (table 1; figure 1).

A total of 431 proteins were significantly differentially abundant (Benjamini-Hochberg adjusted $p<0·05$) between bacterial and viral infections in the SomaScan dataset, with 198 more abundant and 233 proteins less abundant in bacterial infections than in viral infections (figure 2A). In the MS-A dataset, 54 proteins were significantly differentially abundant between bacterial and viral infections with 20 proteins more abundant and 34 proteins less abundant in bacterial infections than in viral infections (figure 2B). In the MS-B dataset, 97 proteins were significantly differentially abundant between bacterial and viral infections with 28 proteins more abundant and 69 proteins less abundant in bacterial infections than in viral infections (figure 2C). 16 proteins were significantly differentially abundant between bacterial and viral infections in all three datasets with

	SomaScan		MS-A		MS-B	
	Definite bacterial (n=48)	Definite viral (n=31)	Definite bacterial (n=74)	Definite viral (n=73)	Definite bacterial (n=75)	Definite viral (n=75)
Study of origin	EUCLIDS	EUCLIDS	EUCLIDS	EUCLIDS	PERFORM	PERFORM
Sample type	Serum	Serum	Plasma	Plasma	Plasma	Plasma
Age (months)	25 (9–77)	11 (4–35)	35 (10–61)	9 (2–23)	57 (12–101)	37 (10–83)
Female	24 (50%)	17 (55%)	47 (64%)	35 (48%)	33 (44%)	33 (44%)
Male	24 (50%)	14 (45%)	27 (37%)	38 (52%)	42 (56%)	42 (56%)
Duration of symptoms (days)	3 (2–7)	3 (1–4)	2 (1–4)	3 (2–7)	2 (1–5)	3 (1–6)
C-reactive protein (mg/L)	124·0 (68·0–226·0)	14·5 (5·0–34·8)	108·0 (39·3–220·8)	17·5 (6·0–26·5)	72·5 (20·7–146·7)	10·3 (3·2–22·0)
Platelets (×10⁹/L)	216 (128–334)	249 (205–376)	228 (158–351)	306 (228–388)	263 (207–356)	274 (199–339)
Lymphocytes (×10⁹/L)	1·8 (1·1–3·6)	2·0 (1·4–3·1)	1·5 (0·6–2·7)	2·2 (1·7–3·1)	1·4 (1·0–3·0)	2·5 (1·6–4·1)
Neutrophils (×10⁹/L)	8·7 (3·6–14·1)	6·3 (2·9–10·2)	4·7 (1·8–12·9)	3·7 (2·7–6·4)	11·2 (6·1–17·3)	4·7 (2·7–7·3)
Monocytes (×10⁹/L)	0·8 (0·4–1·3)	1 (0·7–1·2)	1·5 (0·7–3·0)	0·9 (0·4–1·2)	1·1 (0·6–1·4)	0·8 (0·5–1·1)
Most common causative pathogens or disease	*Neisseria meningitidis* (n=11); *Staphylococcus aureus* (n=8); *Streptococcus pneumoniae* (n=5); *Streptococcus pyogenes* (n=5)	*Enterovirus* (n=6); *Respiratory syncytial virus* (n=5); *Rhinovirus* (n=4)	*Neisseria meningitidis* (n=28); *Streptococcuspneumoniae* (n=13); *Staphylococcus aureus* (n=9); *Escherichia coli* (n=7)	*Respiratory syncytial virus* (n=22); *Enterovirus* (n=13); *Rhinovirus* (n=9); *Influenza virus* (n=6)	*Escherichia coli* (n=21); *Streptococcus pyogenes* (n=14); *Staphylococcus aureus* (n=9); *Neisseria meningitidis* (n=5)	*Influenza virus* (n=19); *Rhinovirus* (n=10); *Adenovirus* (n=9); Epstein–Barr virus (n=7)

Data are n (%) and median (IQR). Percentages might not sum to 100 because of rounding.

Table 1: **Clinical and laboratory features of patients whose samples were included in the discovery of novel protein biomarkers (ie, the high-throughput screening phase)**

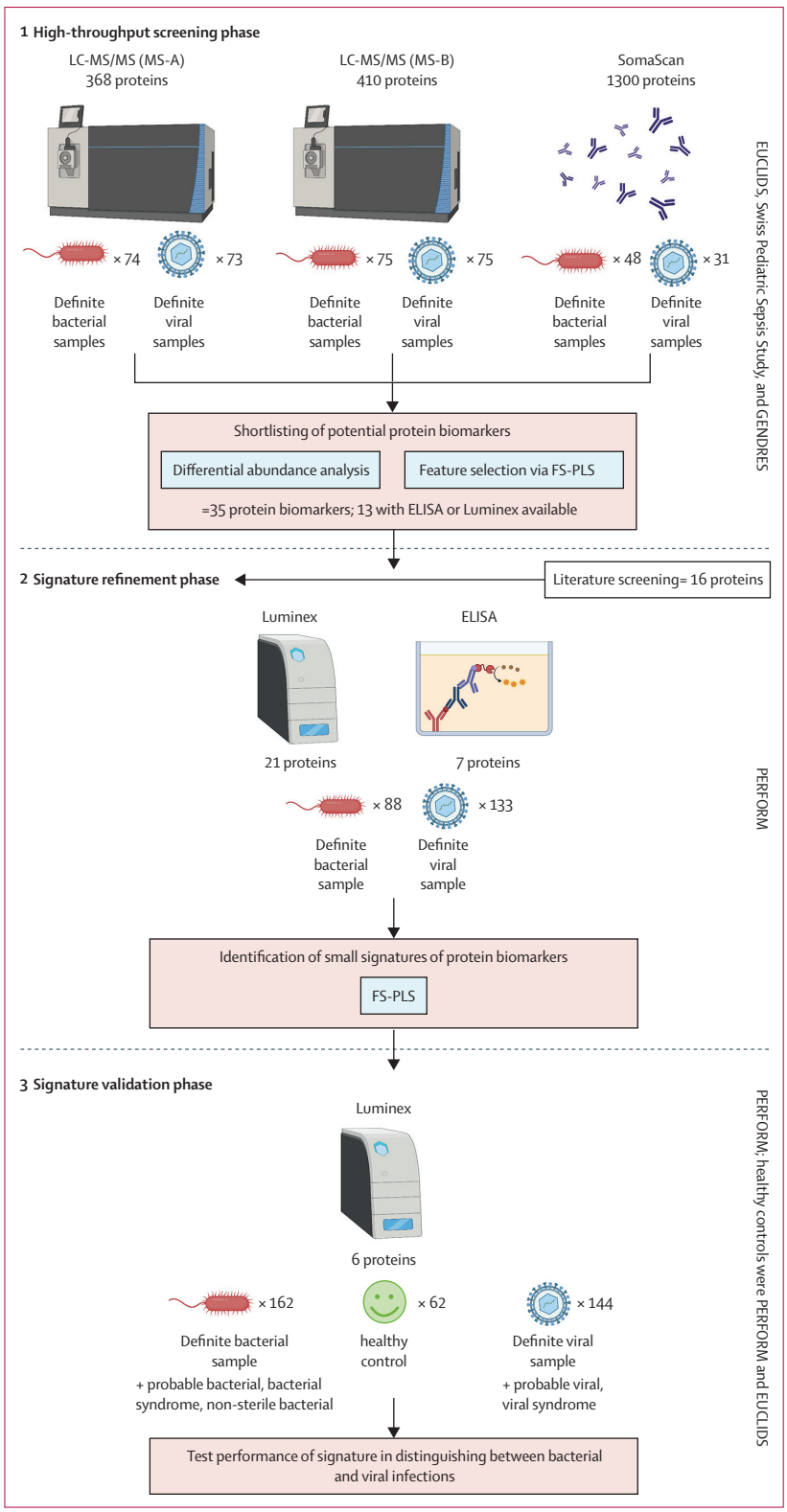

Figure 1: **Procedures flowchart**
FS-PLS=forward selection-partial least squares. LC-MS/MS=liquid chromatography-tandem mass spectrometry. Figure created with BioRender.com.

concordant log-fold change directions, and these proteins were added to the shortlist of potential protein biomarker candidates (table 2).

When iterative FS-PLS (with 100 iterations) was applied to the SomaScan dataset, a three-protein signature was selected the most frequently (five of 100 iterations), composed of ISG15 (robustness 0·92), TIMP metallopeptidase inhibitor 1 (TIMP1; 0·42), and UL16 binding protein 3 (ULBP3; 0·05). When iterative FS-PLS was applied to the MS-A dataset, a four-protein signature was selected the most frequently (five of 100 iterations), composed of liposaccharide binding protein (LBP; robustness 0·55), clusterin (CLUS; 0·38), apolipoprotein H (APOH; 0·32), and histidine-rich glycoprotein (HRG; 0·08). When iterative FS-PLS was applied to the MS-B dataset, a five-protein signature was selected the most frequently (seven of 100 iterations), composed of antithrombin 3 (AT3; robustness 0·95), ceruloplasmin (CERU; 0·49), secreted phosphoprotein 24 (SPP24; 0·29), apolipoprotein C1 (APOC1; 0·17) and α1-antichymotrypsin (AACT; 0·09).

A total of 35 protein biomarker candidates were identified in the high-throughput screening phase and considered for quantification in the signature refinement phase using an independent set of samples, including 18 proteins more abundant in bacterial infections, and 17 proteins more abundant in viral infections (table 2, appendix p 12). A total of 13 protein targets had commercial Luminex or ELISA assays available, and were taken forward to the signature refinement phase, including five that increased in bacterial infections and eight that increased in viral infections.

Literature searches were done in parallel to the hypothesis-free high-throughput screening phase to identify further protein biomarker candidates. A total of 16 potential protein biomarkers were identified from the literature that also had commercial Luminex or ELISA assays available (appendix p 7), including ten that were found to increase and six that were found to decrease in bacterial infections.

The concentrations of the 13 proteins identified in the high-throughput screening phase as potential biomarkers for distinguishing between definite bacterial and definite viral infections with commercially available ELISA or Luminex immunoassays were evaluated in addition to the 16 proteins identified from the literature (appendix pp 7–8, 13). 88 definite bacterial and 113 definite viral infection samples were obtained from patients recruited into the PERFORM study for this evaluation in the signature refinement phase (table 3). Of the 13 proteins from the high-throughput screening phase, 10 were significantly different between bacterial and viral infections and of the 16 proteins derived from the literature, 11 were significantly different between bacterial and viral infections when concentrations were compared using Mann-Whitney U tests (appendix pp 8, 13).

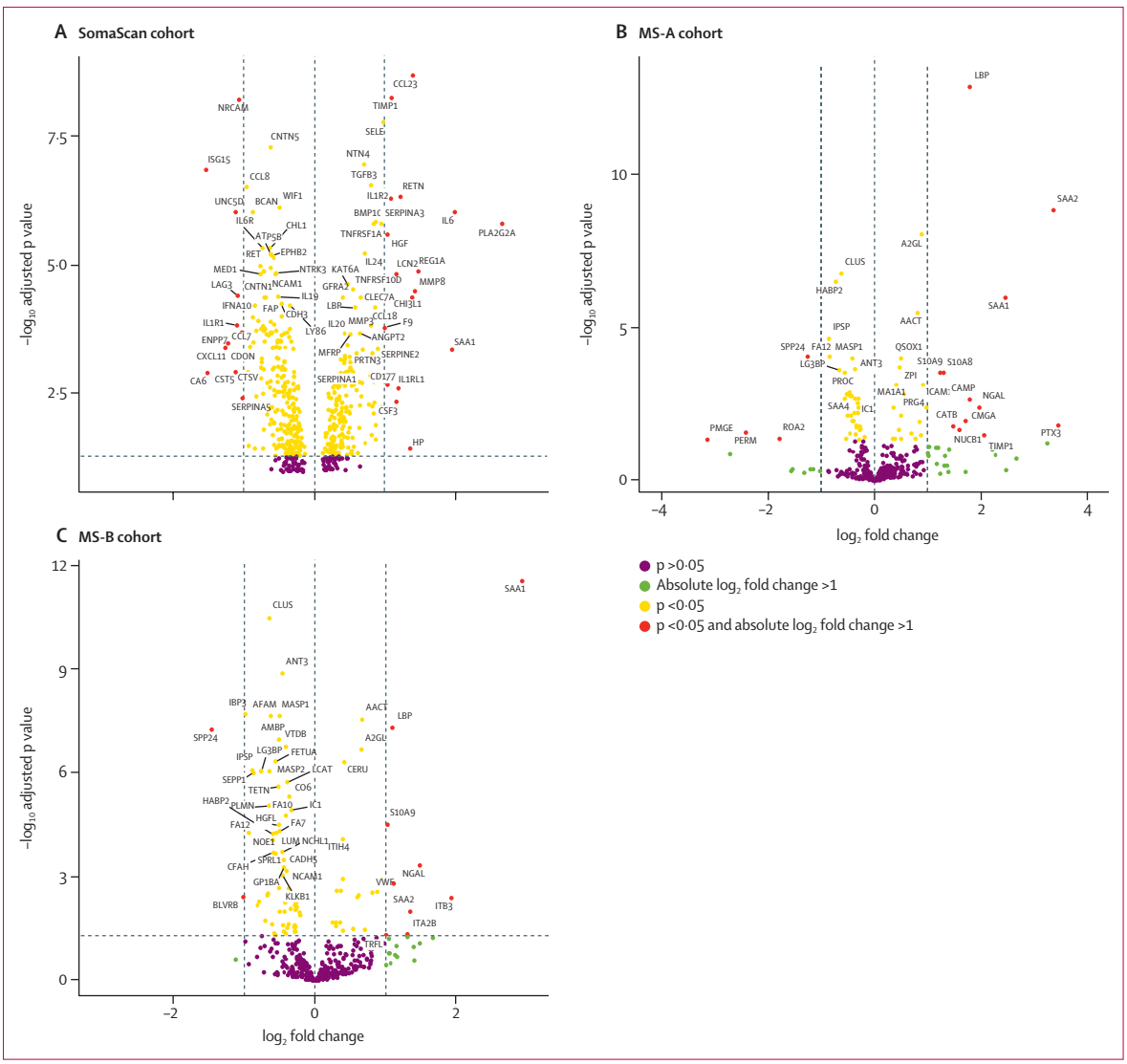

Figure 2: **Volcano plots for differential abundance analysis for each dataset**
Volcano plots show the $\log_2$ fold-change values and the $-\log_{10}$ Benjamini-Hochberg adjusted p values for proteins in the (A) SomaScan, (B) MS-A (C) and MS-B cohorts for models contrasting definite bacterial and definite viral samples. Axis scales vary between A, B, and C for readability.

Iterative FS-PLS was applied to narrow down the list of 29 protein candidates and identify small protein signatures for distinguishing definite bacterial from definite viral infection, with and without the proteins measured using ELISA. When FS-PLS was applied to the 21 proteins measured using Luminex assays, the most frequently selected signature was a five-protein signature composed of NCAM1, IL18, SELE, NGAL, and IFN-γ (the Luminex signature), which was selected in seven of 25 iterations. When FS-PLS was applied to the 28 proteins measured by ELISA assays (except C-reactive protein since it was used in the classification of patients with definite viral infection) in addition to the Luminex proteins, the most frequently selected signature was a five-protein signature composed of SELE, IL18, NCAM2, SAA1, ANGPT2 (the Luminex–ELISA signature), which was selected in seven of 25 iterations. The ROC curves for the Luminex and the Luminex–ELISA signatures (SELE, IL18, NCAM1, SAA1, ANGPT2) were compared using the roc.test function from the pROC package and no significant difference was identified (p=0·84). Despite this result, the Luminex signature had a higher overall AUC (Luminex 89·1% *vs* Luminex–ELISA 88·7%; figure 3A; appendix p 9), higher maximum sensitivity (90·4% *vs* 89·7%), specificity (67% *vs* 64·4%), and partial AUC when specificity was limited to 90–100% (6·1% *vs* 5·3%; appendix p 9) and 95–100% (2·4% *vs* 2·3%; appendix p 9). Furthermore, the Luminex signature led to fewer misclassifications of definite viral infections than the Luminex–ELISA signature (appendix p 9). The

	Protein name	Uniprot	Infection type protein is elevated in	Reasons for inclusion
A2GL	Leucine-rich α-2-glycoprotein	P02750	Bacterial	MS-A Limma top five, MS-A most robust (n=2)
AACT	α-1-antichymotrypsin	P01011	Bacterial	Significantly differentially abundant in all three datasets, MS-B top signature (n=2)
AFM	Afamin	P43652	Viral	MS-B top five significantly differentially abundant (n=1)
AT3	Antithrombin-III	P01008	Viral	MS-B top signature, MS-B top five significantly differentially abundant, MS-B most robust (n=3)
APOC1	Apolipoprotein C-I	P02654	Bacterial	MS-B top signature (n=1)
APOH	β-2-glycoprotein 1	P02749	Bacterial	MS-A top signature, MS-A most robust (n=2)
CERU	Ceruloplasmin	P00450	Bacterial	MS-B top signature, MS-B most robust (n=2)
CLUS	Clusterin	P10909	Viral	Significantly differentially abundant in all three datasets, MS-A top signature, MS-A top five significantly differentially abundant, MS-B top five significantly differentially abundant, MS-A most robust (n=5)
CNTN5	Contactin-5	O94779	Viral	SomaScan top five significantly differentially abundant (n=1)
CO7	Complement component C7	P10643	Viral	Significantly differentially abundant in all three datasets (n=1)
FA5	Coagulation factor V	P12259	Viral	Significantly differentially abundant in all three datasets (n=1)
FBLN3	EGF-containing fibulin-like extracellular matrix protein 1	Q12805	Bacterial	Significantly differentially abundant in all three datasets (n=1)
FETUA	α-2-HS-glycoprotein	P02765	Viral	Significantly differentially abundant in all three datasets (n=1)
HRG	Histidine-rich glycoprotein	P04196	Bacterial	MS-A top signature (n=1)
IGFBP3	Insulin-like growth factor-binding protein 3	P17936	Viral	MS-B top five significantly differentially abundant (n=1)
IPSP	Plasma serine protease inhibitor	P05154	Viral	Significantly differentially abundant in all three datasets (n=1)
ISG15	Ubiquitin-like protein ISG15	P05161	Viral	SomaScan top signature, SomaScan most robust (n=2)
KAIN	Kallistatin	P29622	Viral	MS-B most robust (n=1)
LBP	Lipopolysaccharide-binding protein	P18428	Bacterial	Significantly differentially abundant in all three datasets, MS-A top signature, MS-A top five significantly differentially abundant, SomaScan most robust, MS-A most robust (n=5)
LG3BP	Galectin-3-binding protein	Q08380	Viral	Significantly differentially abundant in all three datasets (n=1)
MASP1	Mannan-binding lectin serine protease 1	P48740	Viral	Significantly differentially abundant in all three datasets, MS-A most robust (n=2)
MASP2	Mannan-binding lectin serine protease 2	O00187	Viral	MS-B most robust (n=1)
MPIF1	C-C motif chemokine 23	P55773	Bacterial	SomaScan top five significantly differentially abundant (n=1)
NCAM1	Neural cell adhesion molecule 1	P13591	Viral	Significantly differentially abundant in all three datasets (n=1)
NGAL	Neutrophil gelatinase-associated lipocalin	P80188	Bacterial	Significantly differentially abundant in all three datasets (n=1)
NRP1	Neuropilin-1	O14786	Bacterial	Significantly differentially abundant in all three datasets (n=1)
PLG	Plasminogen	P00747	Viral	Significantly differentially abundant in all three datasets (n=1)
SAA1	Serum amyloid A-1	P0DJI8	Bacterial	Significantly differentially abundant in all three datasets, MS-A top five significantly differentially abundant, MS-B Limma top five (n=3)
SAA2	Serum amyloid A-2	P0DJI9	Bacterial	MS-A top five significantly differentially abundant (n=1)
SELE	E-Selectin	P16581	Bacterial	SomaScan top five significantly differentially abundant, SomaScan most robust (n=2)
SPP24	Secreted phosphoprotein 24	Q13103	Viral	MS-B top signature, MS-B most robust (n=3)
TIMP1	Metalloproteinase inhibitor 1	P01033	Bacterial	SomaScan top signature, SomaScan top five significantly differentially abundant, SomaScan most robust (n=3)
TNF sR-I	Tumor necrosis factor receptor superfamily member 1A	P19438	Bacterial	SomaScan most robust (n=1)
ULBP3	UL16-binding protein 3	Q9BZM4	Bacterial	SomaScan top signature (n=1)
ZPI	Protein Z-dependent protease inhibitor	Q9UK55	Bacterial	Significantly differentially abundant in all three datasets (n=1)

Ns in parentheses following reasons for inclusion indicate the total number of reasons.

Table 2: **Shortlist of proteins candidates identified from the high-throughput screening phase.**

Luminex signature was taken forward to the signature validation phase over the Luminex–ELISA due to its higher sensitivity which was prioritised due to the implications of missing a severe bacterial infection.

Differential abundance analysis was done, comparing definite bacterial samples with C-reactive protein concentrations of 60 mg/L or less with definite viral samples. Galectin-3-binding protein (LG3BP) was the

	Signature refinement phase		Signature validation phase							
	Definite bacterial (n=88)	Definite viral (n=113)	Definite bacterial (n=162)	Non-sterile definite bacterial (n=31)	Probable bacterial (n=64)	Bacterial syndrome (n=2)	Definite viral (n=144)	Probable viral (n=75)	Viral syndrome (n=12)	Healthy control (n=62)
Age (months)	50 (13–115)	38 (13–76)	53 (12–120)	116 (35–188)	47 (19–101)	71 (47–71)	44 (10–92)	29 (9–64)	27 (10–44)	106 (68–163)
Female	45 (51%)	51 (45%)	73 (45%)	11 (36%)	28 (44%)	0	72 (50%)	31 (41%)	4 (33%)	23 (37%)
Male	43 (49%)	62 (55%)	89 (55%)	20 (65%)	36 (56%)	2 (100%)	72 (50%)	44 (59%)	8 (67%)	39 (63%)
Duration of symptoms (days)	2 (1–4)	3 (1–5)	3 (1–5)	2 (2–7)	3 (2–5)	4 (3–4)	3 (1–5)	2 (1–4)	3 (2–6)	..
C-reactive protein (mg/L)	103·0 (33·2–202·2)	10·3 (4·0–23·8)	89·0 (32·5–175v0)	79·1 (25·0–129·2)	106·7 (53·9–179·0)	7·6 (7·6–7·6)	6·6 (3·0–18·0)	4·0 (2·2–10·8)	8·6 (2·5–30·2)	..
Platelets (×10⁹/L)	264 (166–395)	270 (201–349)	284 (230–396)	295 (215–365)	285 (220–397)	265 (259–270)	250 (192–337)	285 (222–371)	353 (309–407)	..
Lymphocytes (×10⁹/L)	2·1 (1·1–3·5)	2·8 (1·7–3·8)	2·3 (1·2–3·3)	1·8 (1·5–2·6)	2·2 (1·3–3·3)	3·2 (1·9–4·4)	2·3 (1·3–3·8)	3·1 (1·8–4·8)	3·6 (2·8–5·4)	..
Neutrophils (×10⁹/L)	11·4 (5·5–17·6)	5·1 (2·7–6·5)	10·1 (6·7–15·9)	5·6 (4·2–8·8)	11·7 (7·4–16·9)	7·0 (6·9–7·1)	3·6 (2·2–6·3)	3·8 (2·4–6·7)	13·3 (8·5–16·6)	..
Monocytes (×10⁹/L)	1·0 (0·6–1·6)	0·8 (0·5–1·1)	1·2 (0·7–1·6)	0·9 (0·6–1·1)	1·3 (0·8–2·0)	0·8 (0·8–0·8)	0·6 (0·4–1·0)	0·7 (0·4–1·0)	1·7 (1·0–2·0)	..
Causative pathogens or disease	*Escherichia coli* (n=16); *Streptococcus pyogenes* (n=11); *Staphylococcus aureus* (n=11); *Neisseria meningitidis* (n=11); *Streptococcus pneumoniae* (n=9); other* (n=30)	Adenovirus (n=19); influenza virus (n=18); rhinovirus (n=15); respiratory syncytial virus (n=12); enterovirus (n=12); other* (n=37)	*Escherichia coli* (n=70); *Staphylococcus aureus* (n=31); *Streptococcus pneumoniae* (n=23); *Streptococcus pyogenes* (n=11); other* (n=27)	*Salmonella* spp (n=11); *Mycoplasma* spp (n=8); *Campylobacter* spp (n=6); *Mycobacterium tuberculosis* (n=3); other* (n=3)	*Streptococcus* spp (n=10); *Staphylococcus* spp (n=7); *Clostridium difficile* (n=3); *Campylobacter* spp (n=3); other* (n=41)	..	Influenza virus (n=45); Epstein–Barr virus (n=21); Enterovirus (n=19); Respiratory syncytial virus (n=18); other* (n=41)	Rhinovirus (n=7); measles (n=7); enterovirus (n=4); Epstein–Barr virus (n=4); other* (n=53)	Rhinovirus (n=3); respiratory syncytial virus (n=2); adenovirus (n=2); influenza virus (n=2); other* (n=3)	..

Data are n (%) and median (IQR). *Other refers to less frequently identified pathogens.

***Table 3*: Clinical and laboratory features of patients whose samples were included in the signature refinement and signature validation phase**

most significantly differentially abundant protein, with a Benjamini-Hochberg[16] adjusted p value of 0·013 (log_2 fold-change –0·706). LG3BP was taken forward to the signature validation phase in addition to the Luminex signature.

Concentrations of the proteins included in the signature identified in the signature refinement phase were tested on an independent set of plasma samples from PERFORM patients with definite bacterial (n=162) and definite viral (n=144) infections, non-sterile definite bacterial infection (n=31), probable bacterial infection (n=64), bacterial syndrome (n=2), probable viral infection (n=75), viral syndrome (n=12), and healthy controls (n=61; table 3) in the signature validation phase.

The performance of the five-protein Luminex signature (appendix p 9) in classifying definite bacterial and definite viral infection samples was tested. The AUC calculated using the original model weights from the signature refinement phase was 79·7% (95% CI 74·9–84·6; figure 3B; appendix p 9). This result improved to 89·2% (85·7–92·7) when retrained model weights were used.

The measurements of the proteins included in the signature were combined into a single score for each sample—a simple disease risk score (appendix p 4).[19] The direction (ie, whether the proteins are expected to increase or decrease) was identified from the weights calculated by FS-PLS in the signature refinement phase. The AUC using the simple disease risk score was 87·4% (95% CI 83·6–91·2; figure 3B; appendix p 9).

LG3BP was also taken forward from the signature refinement phase for further validation and was combined with the five-protein signature, leading to a slightly improved AUC of 89·3% (95% CI 85·7–92·9; figure 3B) when the simple disease risk score was used. When model weights were retrained with LG3BP included in the signature, the six-protein signature had an AUC of 93·6% (90·9–96·3). The addition of LG3BP led to statistically significant differences between the ROC models for the five-protein and six-protein signatures calculated using retrained model weights ($p=1·5\times10^{-4}$) but not for the models calculated using the simple disease risk score. The addition of LG3BP to the signature led to improvements in specificity over the five-protein signature but not sensitivity. The specificity of the six-protein signature was 89·6% for ROC models with the retrained weights and 85·4% for the simple disease risk score (appendix p 10).

The six-protein signature was used for downstream analyses given its improved specificity in classifying

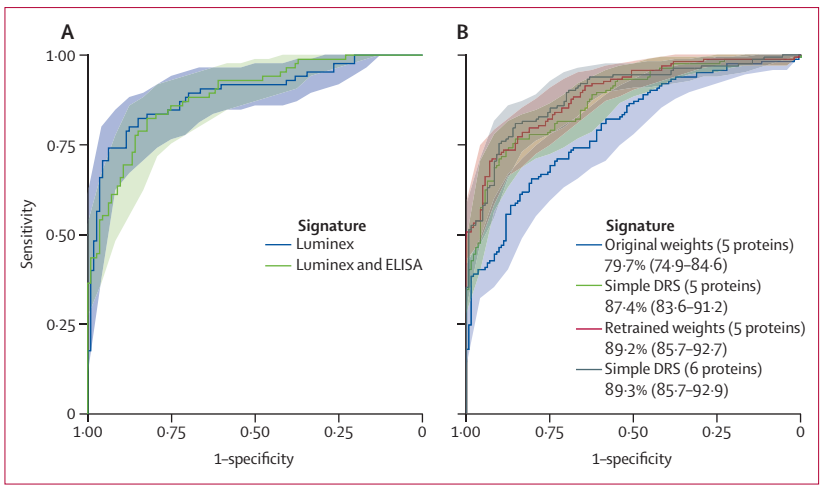

Figure 3: **ROC curves for the signature refinement phase and the signature validation phase**
(A) The ROC curves of the protein signatures identified by running FS-PLS on either Luminex or Luminex and ELISA proteins measured in the signature refinement phase. (B) The ROC curves of the five-protein signature identified in the signature validation phase calculated using original logistic regression weights, retrained general linear model regression weights, the simple DRS for the five-protein signature, and the simple DRS for the six-protein signature. AUC (95% CI) is shown for each category. AUC ROC=area under the curve of the receiver operating characteristic. DRS=disease risk score. FS-PLS=forward selection-partial least squares.

definite viral samples (appendix p 10). The six-protein signature was applied to the other phenotypic groups to identify whether they would be classified as bacterial or viral (appendix pp 5–6, 14) and was compared with the three-protein signature identified by Oved and colleagues (MeMed BV).[9] The six-protein signature was compared with the three-protein signature with and without C-reactive protein and, in both instances, the six-protein signature outperformed the three-protein signature with statistically significant differences in AUC (models including C-reactive protein $p=5{\cdot}92\times10^{-4}$; models excluding C-reactive protein $p=7{\cdot}79\times10^{-8}$; appendix pp 5–6, 15).

Discussion

We did a multi-platform, multi-cohort study to identify and subsequently validate protein biomarkers for differentiating between bacterial and viral infections in children who are febrile. We have identified a five-protein signature with AUCs ranging between 79·9% and 89·2% and specificity of 76·5% in distinguishing between definite bacterial and viral infections. The performance of the signature improved following the addition of LG3BP to the signature, with increases to the specificity (up to 89·6%) and AUC. Using a simple performance evaluation metric—the simple disease risk score—the combination of six proteins (ie, SELE, IL18, NCAM1, NGAL, IFN-γ, and LG3BP) had an AUC of 89·3%, which increased to 93·6% following retraining of model weights. However, this performance with retrained model weights is probably an over-estimation of the signature performance as the model weights are likely to be overfitted to the data, so the simple disease risk score performance should be used.

Of the six proteins included in the six-protein signature, three were elevated in bacterial patients (ie, SELE, NGAL, and IFN-γ), and three were elevated in viral patients (ie, IL18, NCAM1, and LG3BP). All proteins in the signature have biological functions relevant to the host response to infection. SELE is a glycoprotein expressed on endothelial cells after activation by IL-1, TNF-α, or bacterial lipopolysaccharide.[20] SELE mediates leukocyte rolling and is involved in neutrophil, monocyte, and T-cell recruitment to inflammatory foci.[20] NGAL has a fundamental role in the control of bacterial infections by preventing iron acquisition through sequestering iron-loaded bacterial siderophores, which provide essential iron nutrients to the bacteria, thus preventing their survival.[21] IFN-γ is a proinflammatory cytokine with crucial roles in innate and adaptive immunity, including a protective role against bacterial infections.[22] IL18 is a proinflammatory cytokine that induces other inflammatory cytokines. IL18 promotes cell activation of Th1 cells and enhances cytotoxic activity of CD8 T cells and natural killer cells.[23] IL18 was reported to be elevated following various viral infections.[24,25] NCAM1 is a glycoprotein with various functions, and has been identified as a potential viral receptor for rabies virus[26] and Zika virus.[27] LG3BP is a soluble scavenger receptor that has been identified as being associated with various viral infections including Dengue virus[28] and HIV.[29]

The combination of C-reactive protein, TRAIL, and IP10 has been reported as being a promising protein signature for diagnosing febrile children, with a sensitivity of 93·7% and specificity of 94·2%.[10] C-reactive protein was used in the initial classification of the patients with definite viral infection reported here, as per the PERFORM phenotyping algorithm (appendix p 11), meaning direct comparison is challenging. Despite these challenges, the six-protein signature presented here outperformed the Oved and colleagues[9] three-protein signature both with and without the inclusion of C-reactive protein, leading to statistically significant improvements in performance. C-reactive protein is an imperfect biomarker, with elevated C-reactive protein in various other infectious and inflammatory conditions, including SARS-CoV-2, influenza, and severe adenovirus.[5,30] The lower performance of IP10 and TRAIL in classifying the bacterial and viral samples used in our analyses could reflect differences in the patient populations used between our study and Papan and colleagues' study,[10] and different protein detection methods and antibody clones in the MeMed BV test compared with the Luminex assays presented here.

This study is not without limitations. First, co-infection of bacterial and viral pathogens can occur, meaning that some patients with definite bacterial infection might have also had viruses present. Despite this possibility, the six-protein signature can accurately classify 90% of definite bacterial and 82% of definite viral patients,

meaning that it is expected to be robust to co-infections. Second, some promising biomarker candidates identified in the high-throughput screening phase could not be validated due to an absence of commercially available assays. Third, as the literature review was completed on Dec 01, 2017, the proteins identified in this process only reflect those detailed in the literature before this date. Fourth, this study was composed of populations from primarily high-income settings from hospital settings across Europe, meaning that lower-middle-income settings have not been represented and further validation would be required in these settings. Finally, the spectrum of disease-causing pathogens differs between the cohorts included in the SomaScan and MS-A discovery datasets and the other datasets (MS-B, signature refinement phase, signature validation phase). Despite these differences, the six-protein signature performs well in all datasets, indicating that it is robust to pathogen type. The signature should, however, be validated in a further external cohort to ensure its performance is not specific to the patient cohorts presented here.

Through a rigorous multi-stage study using multiple patient cohorts and platforms, we have discovered and subsequently validated various protein biomarkers, resulting in a six-protein signature that can accurately distinguish between definite bacterial and definite viral infections in children who are febrile. This six-protein signature could be developed into a blood-based rapid point-of-care diagnostic test for distinguishing between bacterial and viral infections in children who are febrile, for example as a rule-out test for establishing who does not need antibiotics. Important next steps would be to identify the optimal way to combine these proteins using a rapid protein quantification platform.

Contributors

HRJ and MK prepared the original manuscript draft. All authors had access to all data in the study. Formal analysis was done by HRJ. The method was developed by HRJ, LJMC, RF, and MWT and software was developed and maintained by HRJ, AJM, and LJMC. JZ, MHJ, TWK, and MIdJ accessed and verified the data. Laboratory work was done by JZ, SM, MSH, RF, HH, MHJ, AMT, and JG. Patients were recruited into any of PERFORM, EUCLIDS, GENDRES or the Swiss Sepsis Study by PKAA, UvB, EDC, IE, MvdDF, RdG, HAM, MP, LJS, DZ, WZ, LJMC, CC-P, AJC, and JAH. The study was conceptualised by LJMC, MK, ML, TWK, MIdJ, JAH, and FM-T. TD was responsible for data curation and maintenance. Study supervision was done by LJMC, MK, ML, TWK, MIdJ, and JAH. Funding was acquired by EDC, ME, PKAA, UvB, CF, MvdF, RdG, HAM, MP, AJP, LJS, MNT, EU, SY, DZ, VJW, MSH, WZ, LJMC, CC-P, AJC, FM-T, and HRJ. All authors wrote the manuscript and were responsible for the final decision to submit the manuscript for publication.

Declaration of interests

AJP, AJC, MP, SM, MNT, ML, WZ, RdG, and UvB disclose payments made to institution through the PERFORM consortium from EU Horizon 2020 Programme (grant number 668303). AJC discloses funding from National Institute for Health and Care Research (NIHR; grant number NIHR134694), EPSRC (grant number EP/T029005/1), and EU Horizon Europe Programme (grant number 848196) in addition to support from the European Society for Paediatric Infectious Disease and Excellence in Paediatrics Institute for attending or travelling to meetings; a patent for a new diagnostic method for infection in children unrelated to the current study; and an unpaid role as Chair of Committee for Scientific Affairs and Awards, European Society for Paediatric Infectious Disease. AJP discloses funding from The Bill & Melinda Gates Foundation, the Wellcome Trust, Cepi, Medical Research Council, and NIHR to their institution, an Institutional (Oxford University) partnership with AstraZeneca for development of COVID-19 vaccines, consulting fees from Shionogi for a COVID-19 vaccine, and acting as unpaid chair of the UK's Department of Health and Social Care's Joint Committee on Vaccination and Immunisation and an unpaid member of WHO's SAGE until 2022. MWT discloses support from Janssen and Pfizer for attending or travelling to meetings, unpaid contributions to the National Committee on Immunization Practices, Greece and the Scientific Advisory Group of Experts for COVID-19, Greece, and acting as the President of the Hellenic Society for Paediatric Infectious Diseases. FM-T discloses consulting fees from Sanofi, MSD, Moderna, GSK, Biofabri, AstraZeneca, Novovax, Jansenn, and Pfizer as honoraria for lectures and presentations; support from Pfizer, MSD, GSK, and Sanofi for travel expenses and meeting fees; participation on a data safety monitoring board or advisory board for Pfizer and Biofabri; is a member of The European Technical Advisory Group of Experts (ETAGE) with WHO Europe, Coordinator of Spanish Pediatric Critical Trials Network, and Coordinator of WHO Collaborating Centre for Vaccine Safety of Santiago de Compostela; and is principal investigator in randomised controlled trials of Ablynx, Abbot, Seqirus, Sanofi, MSD, Merck, Pfizer, Roche, Regeneron, Jansen, Medimmune, Novavax, Novartis, and GSK. PKAA discloses Sanofi Nirsevimab payment to institution. HRJ, JZ, ML, MK, TWK and MIdJ have filed a patent application for the six-protein signature described here. CF is co-founder and Medical Director of Micropathology. All other authors declare no competing interests.

Data sharing

Normalised proteomic data used in the high-throughput screening phase have been deposited on Zenodo with the following links: SomaScan https://doi.org/10.5281/zenodo.7781290; MS-A https://doi.org/10.5281/zenodo.7801523; MS-B https://doi.org/10.5281/zenodo.7801541. De-identified participant data are provided, including disease group, age (months), and sex. Computational code for all analyses have been uploaded to GitHub in the following repository: https://github.com/PIDBG/PERFORM_proteomics.

Acknowledgments

We thank the patients and their families who took part in the studies that the data presented here originated from. This study was supported by all members of the PERFORM consortium. We acknowledge computational resources and support provided by the Imperial College Research Computing Service (http://doi.org/10.14469/hpc/2232). This study was supported by the European Union's Horizon 2020 Programme (grant numbers 668303 for PERFORM, 279185 for EUCLIDS to ML), by the Imperial Biomedical Research Centre of the National Institute for Health Research (grant numbers 206508/Z/17/Z and MRF-160–0008-ELP-KAFO-C0801 to MK) and by the Wellcome Trust and the Medical Research Foundation (grant number 215214/Z/19/Z to HRJ). FM-T received support from Instituto de Salud Carlos III on behalf of the GENDRES study (grant numbers PI10/00540, PI16/01478, PI19/01039, PI16/01569, PI19/01090, and PI22/00406); consorcio Centro de Investigación Biomédica en Red de Enfermedades Respiratorias (grant number CB21/06/00103); GEN-COVID (grant number IN845D 2020/23), and Grupos de Referencia Competitiva (grant number IIN607A2021/05). This study was supported by the Swiss State Secretariat for Education, Research and Innovation under contract number 15.0331 (to PKAA). The opinions expressed and arguments employed herein do not necessarily reflect the official views of the Swiss Government. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1 Herberg JA, Kaforou M, Wright VJ, et al. Diagnostic test accuracy of a 2-transcript host RNA signature for discriminating bacterial *vs* viral infection in febrile children. *JAMA* 2016; **316:** 835–45.

2 Llor C, Bjerrum L. Antimicrobial resistance: risk associated with antibiotic overuse and initiatives to reduce the problem. *Ther Adv Drug Saf* 2014; **5:** 229–41.

3 Sproston NR, Ashworth JJ. Role of C-reactive protein at sites of inflammation and infection. *Front Immunol* 2018; **9:** 754.

4 Simon L, Gauvin F, Amre DK, Saint-Louis P, Lacroix J. Serum procalcitonin and C-reactive protein levels as markers of bacterial infection: a systematic review and meta-analysis. *Clin Infect Dis* 2004; **39:** 206–17.

5 Liu F, Li L, Xu M, et al. Prognostic value of interleukin-6, C-reactive protein, and procalcitonin in patients with COVID-19. *J Clin Virol* 2020; **127:** 104370.

6 Jean Beltran PM, Federspiel JD, Sheng X, Cristea IM. Proteomics and integrative omic approaches for understanding host-pathogen interactions and infectious diseases. *Mol Syst Biol* 2017; **13:** 922.

7 Zandstra J, Jongerius I, Kuijpers TW. Future biomarkers for infection and inflammation in febrile children. *Front Immunol* 2021; **12:** 631308.

8 Leticia Fernandez-Carballo B, Escadafal C, MacLean E, Kapasi AJ, Dittrich S. Distinguishing bacterial versus non-bacterial causes of febrile illness—a systematic review of host biomarkers. *J Infect* 2021; **82:** 1–10.

9 Oved K, Cohen A, Boico O, et al. A novel host-proteome signature for distinguishing between acute bacterial and viral infections. *PLoS One* 2015; **10:** e0120012.

10 Papan C, Argentiero A, Porwoll M, et al. A host signature based on TRAIL, IP-10, and C-reactive protein for reducing antibiotic overuse in children by differentiating bacterial from viral infections: a prospective, multicentre cohort study. *Clin Microbiol Infect* 2022; **28:** 723–30.

11 Hagedoorn NN, Borensztajn DM, Nijman R, et al. Variation in antibiotic prescription rates in febrile children presenting to emergency departments across Europe (MOFICHE): a multicentre observational study. *PLoS Med* 2020; **17:** e1003208.

12 Nijman RG, Oostenbrink R, Moll HA, et al. A novel framework for phenotyping children with suspected or confirmed infection for future biomarker studies. *Front Pediatr* 2021; **9:** 688272.

13 Lachlan Coin. Fspls. 2022. https://github.com/lachlancoin/fspls.git (accessed Nov 1, 2022).

14 R Core Team. R: a language and environment for statistical computing. 2020. (accessed Jan 1, 2022).

15 Ritchie ME, Phipson B, Wu D, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015; **43:** e47.

16 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B* 1995; **57:** 289–300.

17 McClish DK. Analyzing a portion of the ROC curve. *Med Decis Making* 1989; **9:** 190–95.

18 Youden WJ. Index for rating diagnostic tests. *Cancer* 1950; **3:** 32–35.

19 Kaforou M, Wright VJ, Oni T, et al. Detection of tuberculosis in HIV-infected and -uninfected African adults using whole blood RNA expression signatures: a case-control study. *PLoS Med* 2013; **10:** e1001538.

20 Mantovani A, Dejana E. Endothelium. In: Delves PJ, ed. Encyclopedia of Immunology, 2nd edn. Oxford: Elsevier, 1998: 802–06.

21 Nasioudis D, Witkin SS. Neutrophil gelatinase-associated lipocalin and innate immune responses to bacterial infections. *Med Microbiol Immunol (Berl)* 2015; **204:** 471–79.

22 Czarniecki CW, Sonnenfeld G. Interferon-gamma and resistance to bacterial infections. *Acta Pathol Microbiol Scand Suppl* 1993; **101:** 1–17.

23 Van Der Sluijs KF, Van Elden LJ, Arens R, et al. Enhanced viral clearance in interleukin-18 gene-deficient mice after pulmonary infection with influenza A virus. *Immunology* 2005; **114:** 112–20.

24 Julkunen I, Sareneva T, Pirhonen J, Ronni T, Melén K, Matikainen S. Molecular pathogenesis of influenza A virus infection and virus-induced regulation of cytokine gene expression. *Cytokine Growth Factor Rev* 2001; **12:** 171–80.

25 Pirhonen J, Sareneva T, Kurimoto M, Julkunen I, Matikainen S. Virus infection activates IL-1 β and IL-18 production in human macrophages by a caspase-1-dependent pathway. *J Immunol* 1999; **162:** 7322–29.

26 Thoulouze MI, Lafage M, Schachner M, Hartmann U, Cremer H, Lafon M. The neural cell adhesion molecule is a receptor for rabies virus. *J Virol* 1998; **72:** 7181–90.

27 Srivastava M, Zhang Y, Chen J, et al. Chemical proteomics tracks virus entry and uncovers NCAM1 as Zika virus receptor. *Nat Commun* 2020; **11:** 3896.

28 Liu KT, Liu YH, Chen YH, et al. Serum galectin-9 and galectin-3-binding protein in acute dengue virus infection. *Int J Mol Sci* 2016; **17:** 832.

29 Rodríguez-Gallego E, Tarancón-Diez L, García F, et al. Proteomic profile associated with loss of spontaneous human immunodeficiency virus type 1 elite control. *J Infect Dis* 2019; **219:** 867–76.

30 Appenzeller C, Ammann RA, Duppenthaler A, Gorgievski-Hrisoho M, Aebi C. Serum C-reactive protein in children with adenovirus infection. *Swiss Med Wkly* 2002; **132:** 345–50.