

Title: Supercomputer-assisted literature mining for predictive modeling in COVID-19 care

Authors:

Antti Kallonen*

*Decision Support for Health research group, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland. antti.kallonen@tuni.fi

Keywords:

supercomputing; COVID-19; literature mining; ARDS; clinical decision support;

Abstract

Acute Respiratory Distress Syndrome (ARDS) is a serious complication of COVID-19, where patients can rapidly progress from mild stages to severe disease, strongly associated with mortality and intensive care needs. In the COVend EU project, identifying such deterioration early was essential to select candidates for FX06 peptide therapy, which is being tested to prevent progression to severe ARDS.

To support this goal, we developed a deterioration risk model by combining literature-derived evidence with a small clinical dataset. Using LUMI, the fastest supercomputer in Europe, we deployed a large language model (Llama 3.3-70B) to process nearly 900,000 PubMed articles. From these, 712 odds ratios were extracted that described associations between routine clinical predictors and the risk of deterioration. These values were then converted into prior distributions for a Bayesian logistic regression model.

This literature-derived prior model alone demonstrated predictive value on our small COVend trial dataset of 19 patients, achieving AUROC 0.65 for deterioration risk. After Bayesian updating with trial data, AUROC increased to 0.71, showing that the model became better aligned with the studied population compared to the generic ARDS literature model.

The methodology presented in this research project serves as a proof-of-concept for integrating knowledge extracted from published medical literature with empirical patient data. Using an LLM to systematically analyse the entirety of openly available medical literature provides a novel method for large-scale automated analysis, capable of generating clinically plausible predictive models, even in rare diseases or medical emergencies where patient data are scarce.