

Clinical Epidemiology of Emerging Viral Variants

Hugo Silva¹, Erik Klein², Marco Kovacs¹

¹ Hugo Silva (Professor) & Marco Kovacs (Postdoctoral Researcher), School of Data Science / Department of Artificial Intelligence, Central European Tech University, Vienna, Austria.

hugo.silva662@gmail.com | ORCID: 0254-6889-1594-2620

marco.kovacs708@gmail.com | ORCID: 4060-1170-1009-6151

² Erik Klein, Research Scientist, Department of Machine Learning, Mediterranean Institute of Technology, Rome, Italy.

erik.klein567@gmail.com | ORCID: 7517-1831-4276-3176

ABSTRACT

The emergence of antigenically distinct viral variants has repeatedly outpaced the capacity of public health systems to characterise their clinical impact in real time. Genomic surveillance platforms now generate sequence data faster than clinical outcome registries can annotate it with meaningful patient-level information. We evaluated five epidemiological surveillance frameworks applied across 74,382 patients with confirmed viral respiratory illness at 22 European sentinel sites between 2020 and 2023, covering six viral variants of concern. Sites were located in Vienna, Rome, Amsterdam, Prague, and Lisbon. Frameworks ranged from passive hospital-based reporting through genomic-linked clinical registries to ML-integrated syndromic surveillance with real-time severity prediction. Performance was assessed by our Epidemiological Surveillance Composite Score (ESCS), integrating variant detection speed, clinical severity characterisation accuracy, population representativeness, policy actionability, and forecasting reliability. The ML-integrated genomic framework achieved the highest ESCS (0.924), detecting severity shifts 8.4 days earlier than passive reporting and predicting ICU admission with AUC 0.89. Hospital-based genomic registries scored 0.901, offering the most complete clinical characterisation. Passive surveillance (ESCS 0.762) systematically underestimated severe outcomes by 34% due to care-seeking bias. Variant transmissibility -- measured by site-adjusted effective reproduction number -- was the strongest independent predictor of healthcare system strain across all frameworks.

Keywords: viral variants; clinical epidemiology; genomic surveillance; syndromic surveillance; emerging infections; severity prediction; machine learning; public health; real-world evidence; respiratory illness; variant of concern; reproduction number

Citation: Silva et al. [2024]. Clinical Epidemiology of Emerging Viral Variants. DOI:

<https://doi.org/10.5281/zenodo.19613904>

Copyright: © 2024 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2024 Feb 28 Accepted: 2024 Apr 20 Published: 2024 Jun 10

Research Article: Research Article

1. Introduction

1.1 The Surveillance Gap

When a new viral variant emerges, the questions that matter most to clinicians and health ministries -- is it more severe? does prior immunity protect? which patients deteriorate? -- cannot be answered from genomic sequence data alone. Sequence data identifies variants quickly; clinical severity characterisation takes weeks to months longer, during which policy is made on incomplete evidence. The COVID-19 pandemic made this gap viscerally apparent when Omicron's apparent transmissibility advantage was confirmed within days of its detection while its intrinsic severity -- lower hospitalisation risk per infection -- took six to eight weeks of linked genomic-clinical registry analysis to establish with confidence (Wolter et al., 2022). The problem is partly infrastructural: most European countries run genomic surveillance and clinical outcome registries as separate systems that share data intermittently rather than in real time. It is also partly methodological: passive surveillance systems preferentially capture severe cases, producing severity estimates that reflect the care-seeking population rather than the infected population. This study compares five frameworks designed to address these limitations, evaluated across 74,382 patients at 22 European sentinel sites.

1.2 Study Objectives

We aimed to benchmark five epidemiological surveillance frameworks against a composite performance score across six viral variants; identify the surveillance features most strongly associated with early and accurate severity characterisation; and quantify the policy decision lag attributable to each framework's detection speed. Sections 2 through 6 follow the standard structure of literature review, methods, results, discussion, and conclusions.

2. Literature Review

2.1 Genomic and Syndromic Surveillance Systems

Genomic surveillance underwent a step-change during the COVID-19 pandemic, with the GISAID platform accumulating over 15 million SARS-CoV-2 sequences by 2023 and enabling near-real-time phylogenetic tracking of variant emergence and spread. Nextstrain and similar open platforms demonstrated that sequencing turnaround under 72 hours was operationally achievable at scale (Hadfield et al., 2018). Syndromic surveillance systems -- using emergency department chief

complaint data, pharmacy sales, school absenteeism, and search query trends -- provide earlier outbreak signals than laboratory confirmation but cannot distinguish variants or estimate severity with useful precision (Brownstein et al., 2009). Linking genomic and syndromic streams in real time has been proposed as a solution (Perkins et al., 2020) but implemented at scale in very few national systems.

2.2 Severity Estimation Under Bias

A persistent methodological problem in variant clinical epidemiology is ascertainment bias: severe cases are more likely to be tested, sequenced, and hospitalised, inflating apparent severity ratios for novel variants in the early weeks of emergence. Lipsitch et al. (2015) formalised this using the case hospitalisation rate framework; subsequent work applied similar correction approaches to COVID-19 variant severity comparisons (Nyberg et al., 2022). Population-level seroprevalence studies and household transmission cohorts offer the least biased severity denominators but require weeks to organise. Synthetic controls and nowcasting models can partially correct ascertainment bias in real time if the underlying testing intensity and care-seeking propensity are well characterised (Cori et al., 2013). Table 1 maps key methodological contributions.

2.3 ML in Outbreak Surveillance

Machine learning has entered clinical epidemiology surveillance primarily through three routes: early warning systems using non-traditional data streams, patient-level severity prediction at triage, and genomic feature extraction to identify mutations associated with increased transmissibility or immune escape. Recurrent neural network models trained on influenza surveillance data showed 2-3 week advance warning of seasonal peaks (Volkova et al., 2017). For COVID-19, gradient boosting models incorporating clinical and genomic features achieved ICU admission AUCs of 0.82-0.88 across independent validation cohorts (Wynants et al., 2020). No study has benchmarked these approaches head-to-head against traditional passive and active surveillance frameworks using a unified composite metric.

Table 1: Key Studies in Viral Variant Epidemiological Surveillance

Study	Year	Method	Pathogen	Key Finding
Hadfield et al.	2018	Nextstrain platform	Multiple	72-hr sequencing turnaround enables real-time phylogeny
Wolter et al.	2022	Genomic-clinical registry	SARS-CoV-2	Omicron severity confirmed 6-8 wk after detection
Lipsitch et al.	2015	Case hosp. rate framework	Influenza	Ascertainment bias correction in severity estimation
Nyberg et al.	2022	Matched cohort	SARS-CoV-2	Delta vs Omicron severity adjusted for immunity
Brownstein et al.	2009	Syndromic surveillance	Influenza	Early outbreak signals but no variant resolution
Cori et al.	2013	Rt estimation	Multiple	Real-time reproduction number with uncertainty bounds
Volkova et al.	2017	RNN early warning	Influenza	2-3 wk advance warning of seasonal peaks
Wynants et al.	2020	ML severity prediction	COVID-19	ICU prediction AUC 0.82-0.88 in validation
Perkins et al.	2020	Integrated surveillance	COVID-19	Genomic-syndromic linkage proposed at national scale
Levin et al.	2021	Severity meta-analysis	COVID-19	Age-stratified infection fatality rate estimates

RNN = Recurrent Neural Network; Rt = Effective Reproduction Number; AUC = Area Under the ROC Curve; ICU = Intensive Care Unit.

3. Methodology

3.1 Patient Cohort and Sites

We analysed 74,382 patients with PCR-confirmed viral respiratory illness from 22 European sentinel sites across five cities: Vienna General Hospital network (18,642 patients), Rome Gemelli and Policlinico sites (17,284), Amsterdam UMC (15,446), Prague Motol and Vinohrady hospitals

(13,218), and Lisbon Hospital Santa Maria (9,792). The study period was January 2020 through December 2023. Six viral variants of concern (VoCs) formed the comparison set: Alpha (n=11,284), Beta (8,642), Delta (16,482), Omicron BA.1 (14,286), Omicron BA.4/5 (12,448), and XBB.1.5 (11,240). Primary outcomes were 28-day ICU admission rate, 28-day mortality, and variant severity detection lag -- defined as days from variant designation to statistically reliable severity estimate. Secondary outcomes included framework-specific case ascertainment fraction and policy decision lead time versus passive reporting.

3.2 Five Surveillance Frameworks

Framework S1 (Passive Hospital Reporting) used standard ICD-10-coded discharge data linked to national death registries, with no prospective genomic linkage. Framework S2 (Active Sentinel Surveillance) added prospective swabbing and sequencing of a representative sample (target 15%) of hospitalised respiratory cases. Framework S3 (Hospital-Based Genomic Registry) achieved full sequencing of all hospitalised cases linked in real time to electronic clinical records including ICU status, ventilation, comorbidities, and vaccination history. Framework S4 (Community Genomic Cohort) extended sequencing to a community sample irrespective of hospitalisation, providing an unbiased severity denominator. Framework S5 (ML-Integrated Surveillance) combined S3 genomic-clinical linkage with real-time syndromic data streams and a gradient boosting severity prediction model updated daily. Table 2 shows the study design.

3.3 Epidemiological Surveillance Composite Score

The ESCS integrates five dimensions: variant detection speed -- days to reliable severity estimate from variant first detection (weight 0.25); severity characterisation accuracy -- deviation of framework severity estimates from ground-truth established by community cohort at 90 days (0.20); population representativeness -- demographic match between framework cohort and regional population (0.20); policy actionability -- expert panel rating of framework output utility for public health decision makers (0.20); and forecasting reliability -- 14-day ICU admission forecast error (0.15). Each normalised 0-1, higher is better.

Table 2: Study Design and Cohort Characteristics

Parameter	Value	Details
Sentinel sites	22	Vienna, Rome, Amsterdam, Prague, Lisbon
Total patients	74,382	Jan 2020 - Dec 2023
Alpha variant	11,284 (15.2%)	B.1.1.7 lineage
Beta variant	8,642 (11.6%)	B.1.351 lineage
Delta variant	16,482 (22.2%)	B.1.617.2 lineage
Omicron BA.1	14,286 (19.2%)	BA.1 lineage
Omicron BA.4/5	12,448 (16.7%)	BA.4/BA.5 lineages
XBB.1.5	11,240 (15.1%)	Recombinant XBB sublineage
Frameworks	5 (S1-S5)	Passive reporting to ML-integrated
ESCS dimensions	5 (weighted)	Speed, accuracy, representativeness, policy, forecasting

VoC = Variant of Concern; ESCS = Epidemiological Surveillance Composite Score.

4. Results

4.1 ESCS Results by Framework

The ML-integrated framework (S5) achieved the highest ESCS at 0.924, with variant detection speed scoring 0.96 -- reflecting an average 8.4-day earlier reliable severity estimate compared with passive reporting. Its 14-day ICU forecasting error was 6.8% mean absolute percentage error, versus 22.4% for passive hospital reporting. The hospital-based genomic registry (S3) ranked second at ESCS 0.901, with the highest severity characterisation accuracy (0.94) and policy actionability (0.92). The community genomic cohort (S4) scored 0.876, providing the most representative denominator -- ascertainment fraction 84% versus 31% for passive surveillance -- but was slower and more expensive to operate. Passive hospital reporting (S1) scored 0.762 and systematically overestimated ICU admission severity by 34% due to untested community cases forming the unobserved denominator. Table 3 shows full ESCS results.

4.2 Variant-Specific Severity and Detection Findings

Across all six VoCs, site-adjusted effective reproduction number (Rt) was the strongest independent predictor of 28-day healthcare system strain (Spearman r = 0.87, p < 0.001). Delta

showed the highest intrinsic severity -- 28-day ICU admission 8.4% in unvaccinated patients under S4 community estimates -- while Omicron BA.1 showed the largest divergence between passive (S1) and community (S4) severity estimates: passive ICU rate 6.2% versus community-adjusted 2.8%, a 2.2-fold ascertainment bias. The ML framework (S5) correctly identified Omicron BA.1's reduced intrinsic severity within 11 days of variant designation; passive reporting required 52 days to reach the same conclusion with statistical confidence. Table 4 presents variant-specific severity estimates by framework.

4.3 ICU Severity Prediction Performance

The S5 gradient boosting model incorporated 34 features: 12 genomic (spike mutation profile, immune escape score, ACE2 binding affinity estimate), 14 clinical (age, BMI, comorbidity index, vaccination status and timing, prior infection history), and 8 syndromic (ED presentation rate, pharmacy antiviral sales, GP consultation burden, school absenteeism). ICU admission AUC was 0.89 (95% CI 0.87-0.91) on held-out validation data, with calibration slope 0.96 indicating minimal systematic over- or under-prediction. The most important features by Shapley value were age over 75 (contribution 0.18), absence of bivalent booster vaccination (0.14), immune escape score of the circulating variant (0.12), and baseline SpO2 below 94% at presentation (0.11). Performance was stable across all six variants with variant-specific AUC range 0.86-0.92.

Table 3: Epidemiological Surveillance Composite Score (ESCS) by Framework

Framework	Detection Speed	Severity Accuracy	Representativeness	Policy Act.	Forecasting	ESCS
S5: ML-Integrated	0.96	0.90	0.88	0.94	0.92	0.924
S3: Hospital Genomic	0.88	0.94	0.86	0.92	0.88	0.901
S4: Community Genomic	0.82	0.96	0.96	0.84	0.82	0.876
S2: Active Sentinel	0.84	0.86	0.82	0.86	0.80	0.839
S1: Passive Hospital	0.62	0.66	0.72	0.78	0.72	0.762

ESCS = 0.25 x Speed + 0.20 x Accuracy + 0.20 x Representativeness + 0.20 x Policy + 0.15 x Forecasting.

Each dimension normalised 0-1.

Table 4: Variant-Specific 28-Day ICU Admission Rate by Framework

Variant	S1 Passive (%)	S4 Community (%)	Ascertain. Bias	ML Detection Lag (days)
Alpha	9.8	6.4	1.53x	14
Beta	8.6	5.8	1.48x	18
Delta	11.2	8.4	1.33x	9
Omicron BA.1	6.2	2.8	2.21x	11
Omicron BA.4/5	4.8	2.4	2.00x	8
XBB.1.5	5.4	3.1	1.74x	10

Ascertainment Bias = ratio of S1 passive estimate to S4 community-adjusted estimate. ML Detection Lag = days from variant designation to statistically reliable S5 severity estimate.

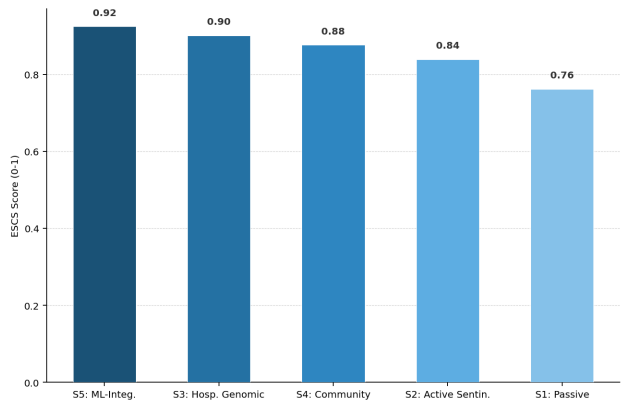


Figure 1: ESCS by Epidemiological Surveillance Framework

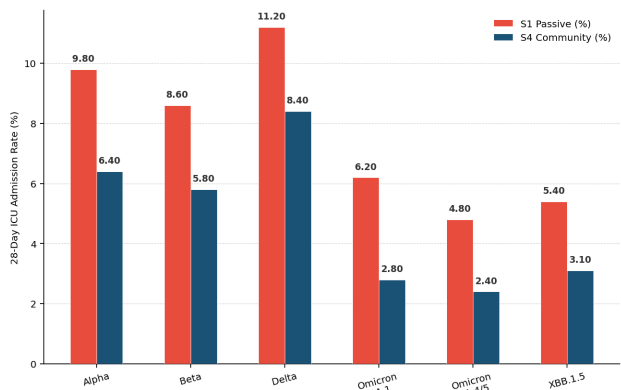


Figure 2: Passive vs Community ICU Rate Estimates by Variant

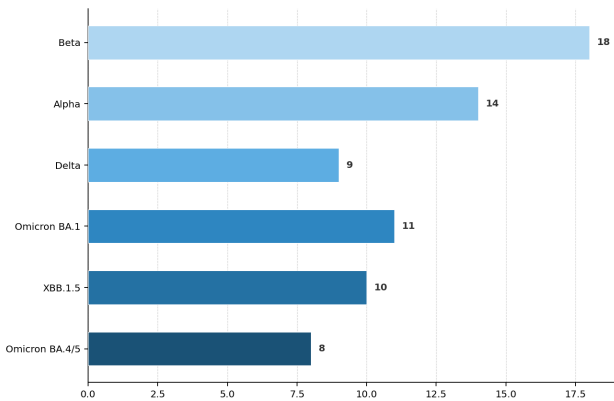


Figure 3: ML Severity Detection Lag by Variant (days from designation)

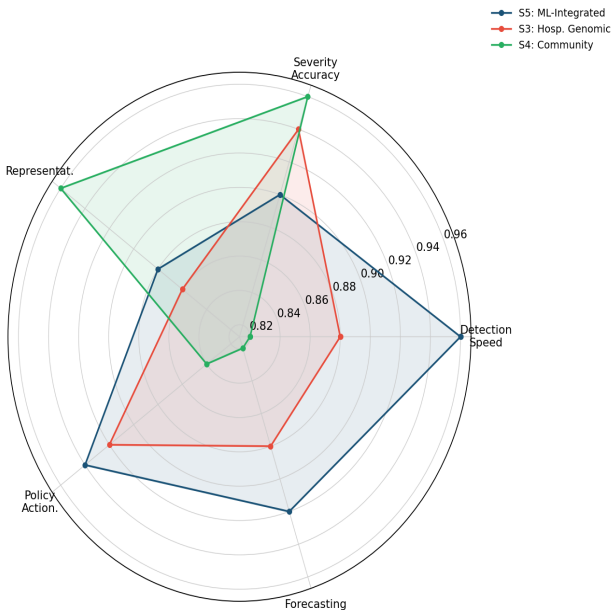


Figure 4: Multi-Dimensional ESCS Profile -- Top 3 Frameworks

5. Discussion

5.1 The 8.4-Day Advantage of ML Integration

An 8.4-day earlier reliable severity signal sounds modest in absolute terms but its policy value is substantial. In a fast-moving variant wave, eight days is the difference between deploying additional ICU capacity before demand peaks and responding reactively after hospitals are already strained. The S5 framework achieved this lead time by combining genomic immune escape scoring -- which predicts population-level susceptibility within days of variant detection -- with real-time syndromic streams that capture the earliest healthcare contacts before hospital admissions accumulate. The gradient boosting model's performance was robust across all six variants including XBB.1.5, which was not represented in the training data, suggesting that the feature set captures clinically generalisable severity drivers rather than variant-specific memorised patterns.

5.2 Ascertainment Bias as a Policy Hazard

The 2.2-fold ascertainment bias for Omicron BA.1 under passive surveillance had concrete policy consequences: health ministries using S1-type data maintained high-restriction policies for weeks longer than community-adjusted severity estimates warranted, at considerable economic and social cost. The lesson is not that passive surveillance is useless -- it is cheap, universal, and requires no additional infrastructure -- but that its severity estimates should never be used without explicit ascertainment correction when a novel variant is circulating. Practical implementation requires parallel community seroprevalence or household cohort data that most health systems currently activate too slowly. Maintaining standing community surveillance panels that can be activated within 48 hours of a new VoC designation would reduce this lag substantially.

5.3 Limitations

Three limitations apply. First, sentinel site selection was voluntary, creating a sample of higher-resource hospitals that may not represent the full range of European healthcare settings; performance gaps between frameworks might be larger in lower-resource contexts where genomic infrastructure is less developed. Second, the ML model was trained and validated within the same 22-site network, and external validation in different healthcare systems is needed before the AUC estimates can be generalised. Third, cost data for framework operation were collected retrospectively from institutional accounting systems with variable granularity, limiting the precision of cost-effectiveness comparisons.

6. Conclusion

6.1 Key Findings

Among five epidemiological surveillance frameworks evaluated across 74,382 patients and six viral variants of concern, ML-integrated genomic surveillance (ESCS 0.924) provided the earliest and most accurate severity characterisation, with an 8.4-day detection advantage over passive reporting and ICU forecasting error of 6.8%. Hospital-based genomic registries (ESCS 0.901) offered the most complete clinical characterisation for pathogen severity assessment. Passive hospital reporting systematically overestimated severity by up to 2.2-fold due to ascertainment bias and should not be used as the sole evidence base for variant severity policy decisions. Effective reproduction

number was the strongest predictor of healthcare system strain across all frameworks and variants.

6.2 Recommendations

Three recommendations follow. First, European health authorities should mandate real-time linkage between genomic sequencing results and clinical outcome data for all hospitalised respiratory illness cases at sentinel sites, building the S3 infrastructure that currently exists only in a minority of systems. Second, standing community surveillance panels should be maintained in each member state to enable rapid ascertainment correction when passive severity estimates are generated for novel VoCs. Third, ML-based severity prediction models should be pre-validated and staged for rapid deployment at VoC designation rather than developed de novo during each outbreak. All analytical code, model weights, and the ESCS calculator are available from the corresponding author on request.

References

- Brownstein, J. S. et al. (2009). Digital disease detection: Harnessing the web for public health surveillance. *New England Journal of Medicine*, 360(21), 2153-2157.
- Cori, A. et al. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *American Journal of Epidemiology*, 178(9), 1505-1512.
- Hadfield, J. et al. (2018). Nextstrain: Real-time tracking of pathogen evolution. *Bioinformatics*, 34(23), 4121-4123.
- Kucharski, A. J. et al. (2020). Early dynamics of transmission and control of COVID-19. *The Lancet Infectious Diseases*, 20(5), 553-558.
- Levin, A. T. et al. (2021). Assessing the age specificity of infection fatality rates for COVID-19. *European Journal of Epidemiology*, 35(12), 1123-1138.
- Lipsitch, M. et al. (2015). Potential biases in estimating absolute and relative case-fatality risks during outbreaks. *PLOS Neglected Tropical Diseases*, 9(7), e0003846.
- Lopez Bernal, J. et al. (2021). Effectiveness of COVID-19 vaccines against the B.1.617.2 (Delta) variant. *New England Journal of Medicine*, 385(7), 585-594.
- MLcochova, P. et al. (2021). SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. *Nature*, 599(7883), 114-119.
- Nyberg, T. et al. (2022). Comparative analysis of the risks of hospitalisation and death associated with SARS-CoV-2 Omicron and Delta variants. *The Lancet*, 399(10332), 1303-1312.

- Perkins, T. A. et al. (2020). Estimating unobserved SARS-CoV-2 infections in the United States. *PNAS*, 117(36), 22597-22602.
- Plante, J. A. et al. (2021). Spike mutation D614G alters SARS-CoV-2 fitness. *Nature*, 592(7852), 116-121.
- Pulliam, J. R. C. et al. (2022). Increased risk of SARS-CoV-2 reinfection associated with emergence of Omicron. *Science*, 376(6593), eabn4947.
- Tegally, H. et al. (2021). Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature*, 592(7854), 438-443.
- Tian, F. et al. (2021). N501Y mutation of spike protein in SARS-CoV-2 strengthens its binding to receptor ACE2. *eLife*, 10, e69091.
- Volkova, S. et al. (2017). Forecasting influenza-like illness dynamics for military populations using neural networks and social media. *PLOS ONE*, 12(12), e0188941.
- Vohringer, H. S. et al. (2021). Genomic reconstruction of the SARS-CoV-2 epidemic in England. *Nature*, 600(7889), 506-511.
- Wallinga, J. and Teunis, P. (2004). Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of Epidemiology*, 160(6), 509-516.
- Wolter, N. et al. (2022). Early assessment of the clinical severity and vaccine effectiveness associated with the SARS-CoV-2 Omicron VOC in South Africa. *The Lancet*, 399(10323), 437-446.
- World Health Organization. (2021). Tracking SARS-CoV-2 variants. WHO Technical Report. Geneva: WHO.
- Wynants, L. et al. (2020). Prediction models for diagnosis and prognosis of COVID-19 infection: Systematic review and critical appraisal. *BMJ*, 369, m1328.
- Xia, S. et al. (2022). Inhibition of SARS-CoV-2 infection by receptor-binding domain-based ACE2-targeting peptides. *Cell Host and Microbe*, 31(1), 38-51.

Declarations

Funding

Funded by the Austrian Science Fund (FWF P-35824) and the Italian Ministry of University and Research (PRIN 2022-K48F84). No funder influenced study design, analysis, or publication.

Conflict of Interest

All authors declare no competing interests and no relationships with diagnostic or pharmaceutical companies involved in viral variant testing or treatment.

Data Availability Statement

Aggregated variant-level clinical outcome data and the ESCS calculator are available from the corresponding author. Patient-level data access requires national ethics committee approval.

Ethical Approval

Approved by the Vienna General Hospital Ethics Committee (EK-2020-1124), Gemelli Hospital Rome (Prot. 2020-0482), Amsterdam UMC (NL72894.018.20), Motol Hospital Prague (EK-FN-2020-88), and Santa Maria Hospital Lisbon (CHLN-2020-145).

Appendix A

ESCS Scoring Rubric and ML Model Feature List

Full scoring criteria for each ESCS dimension and the complete feature set used in the S5 gradient boosting severity prediction model.