Categorising pandemic potential pathogens

Supplementary information

1. Review

1.1 Search strategy

The primary search was conducted in PubMed to identify systematic reviews addressing these parameters for the 19 pathogens selected. The search strategy, outlined in **Supplementary Table S1**, combined pathogen-specific terms with the keywords related to the epidemiological parameters of interest and “systematic review”. Literature published up to July 5, 2024, was included, except for Nipah virus, for which the search extended to September 24, 2024. For SARS and MERS, the terms “NOT SARS-CoV-2 OR COVID-19” were added to exclude irrelevant studies focused on the COVID-19 pandemic. This stage aimed to explore the breadth of existing systematic reviews and identify parameters requiring further targeted searches.

Supplementary Table S1. Search strategies

Pathogen	Parameter	Review
Crimean–Congo hemorrhagic fever OR CCHF OR Crimean Hemorrhagic Fever OR Congo Hemorrhagic Fever	basic reproductive number OR R0 OR basic reproduction number OR basic reproduction ratio OR basic reproductive rate	Systematic review
Ebola OR Ebola virus OR Ebola virus disease OR EBOV OR EVD	Incubation period OR incubation	
Marburg OR Marburg virus OR MARV OR Marburg virus disease OR MVD	Overdispersion OR dispersion parameter	
Lassa OR Lassa fever OR Lassa hemorrhagic fever OR Lassa virus	Latent period OR latency period OR pre-infectious period	
Middle East respiratory syndrome–related coronavirus OR MERS-CoV OR Middle East respiratory syndrome OR MERS	Infectious period	
Severe acute respiratory syndrome OR SARS OR SARS-CoV OR severe acute respiratory syndrome coronavirus	Serial interval	
Rift Valley fever OR RVF	Case fatality ratio OR Case fatality rate OR case-fatality risk OR CFR	
(Zika fever OR Zika virus disease OR Zika OR Zika virus OR ZIKV)	infection fatality ratio OR infection fatality rate OR IFR	
(Nipah virus OR Nipah OR NiV)		
H1N1 influenza A virus OR A/H1N1 OR H1N1 OR Spanish flu OR 1918 influenza pandemic		
Influenza A virus subtype H2N2 OR A/H2N2 OR H2N2 OR 1957–1958		

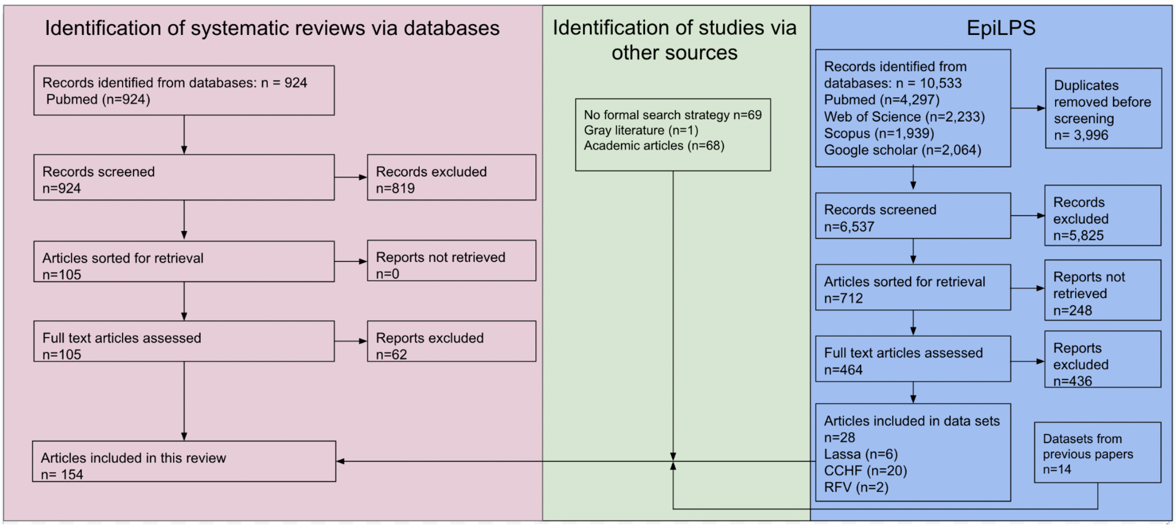
influenza pandemic OR Asian Flu		
Influenza A virus subtype H3N2 OR A/H3N2 OR H3N2 OR Hong Kong Flu OR 1968 flu pandemic		
2009 swine flu OR Pandemic H1N1/09 virus OR 2009 swine flu pandemic		
H5N1 OR Influenza A virus subtype H5N1 OR A/H5N1		
(COVID-19 OR SARS-CoV-2 OR 2019-nCoV OR "coronavirus disease 2019" OR "severe acute respiratory syndrome coronavirus 2") AND (wild type OR ancestral variant OR Wuhan variant)		
(COVID-19 OR SARS-CoV-2 OR 2019-nCoV OR "coronavirus disease 2019" OR "severe acute respiratory syndrome coronavirus 2") AND (Alpha OR Alpha variant OR B.1.1.7)		
(COVID-19 OR SARS-CoV-2 OR 2019-nCoV OR "coronavirus disease 2019" OR "severe acute respiratory syndrome coronavirus 2") AND (Delta OR Delta variant OR B.1.617.2)		
(COVID-19 OR SARS-CoV-2 OR 2019-nCoV OR "coronavirus disease 2019" OR "severe acute respiratory syndrome coronavirus 2") AND (Omicron OR B.1.1.529 OR Omicron variant)		
Mpox OR monkeypox OR Monkeypox virus OR MPV		

Following the database search, titles and abstracts were screened to identify relevant systematic reviews. Studies were included if they provided data on at least one of the specified parameters. Reviews that were not systematic or lacked relevant data were excluded. Full texts of the selected articles were then reviewed to extract data, which included summary statistics, contextual information and probability distributions. All extracted data were documented in a centralised data extraction sheet.

To supplement the systematic reviews, additional searches were conducted to address further gaps in the data. These supplementary searches included narrative and rapid reviews, modeling studies, epidemiological investigations, and human infection studies. Due to the vast volume of potential information, no formal search strategy was applied in this stage. Sources for this search include PubMed, Google Scholar, connected papers [1] (to identify prior and derivative works) and liaising with other academics to identify important papers as well as works in preprint. Data from these supplementary sources were extracted in the same standardised format and integrated into the central data sheet. This search was conducted up to January 2025.

The extracted parameters were categorised based on the source and nature of the data. Parameters extracted from modelling studies have been placed into three categories where applicable: estimated values, where the article has estimated parameters directly from empirical data; referenced values, where the parameter has been taken from other articles; assumed values, where the parameter has been assumed for modelling purposes and has been based on opinion or other data sources.

1.2 Literature search results


Identification of systematic reviews via databases
Records identified from databases: n = 924
Pubmed (n=924)
Records screened
n=924
Records excluded
n=819
Articles sorted for retrieval
n=105
Reports not retrieved
n=0
Full text articles assessed
n=105
Reports excluded
n=62
Articles included in this review
n= 154
Identification of studies via other sources
No formal search strategy n=69
Gray literature (n=1)
Academic articles (n=68)
EpiLPS
Records identified from databases: n = 10,533
Pubmed (n=4,297)
Web of Science (n=2,233)
Scopus (n=1,939)
Google scholar (n=2,064)
Duplicates removed before screening
n= 3,996
Records screened
n=6,537
Records excluded
n=5,825
Articles sorted for retrieval
n=712
Reports not retrieved
n=248
Full text articles assessed
n=464
Reports excluded
n=436
Articles included in data sets
n=28
Lassa (n=6)
CCHF (n=20)
RFV (n=2)
Datasets from previous papers
n=14

Supplementary Figure S1. Flowchart illustrating the review process.

2. Parameter estimation

2.1 Incubation period estimation

To complement the analyses derived from existing literature, the EpiLPS package [2–4] was used to estimate the incubation period of different pathogens when publicly available data permitted. EpiLPS relies on the methodology of Gressani et al. [3] to estimate the incubation distribution by using a flexible semi-parametric Bayesian approach based on Laplacian-P-splines.

Where data was not publicly available, it was collected on previous outbreaks using outbreak reports. For pathogens such as Lassa fever, Crimean-Congo haemorrhagic fever (CCHF), and Rift Valley fever (RVF), searches were conducted in May 2023 using the databases PubMed, Web of Science, Scopus, and Google Scholar. The search strategy combined pathogen-specific terms with keywords such as “community transmission,” “nosocomial transmission,” “reservoir transmission,” and “outbreak” or “cluster.” Relevant reports were identified based on their inclusion of exposure timelines and corresponding symptom onset data, which were subsequently used to estimate incubation periods.

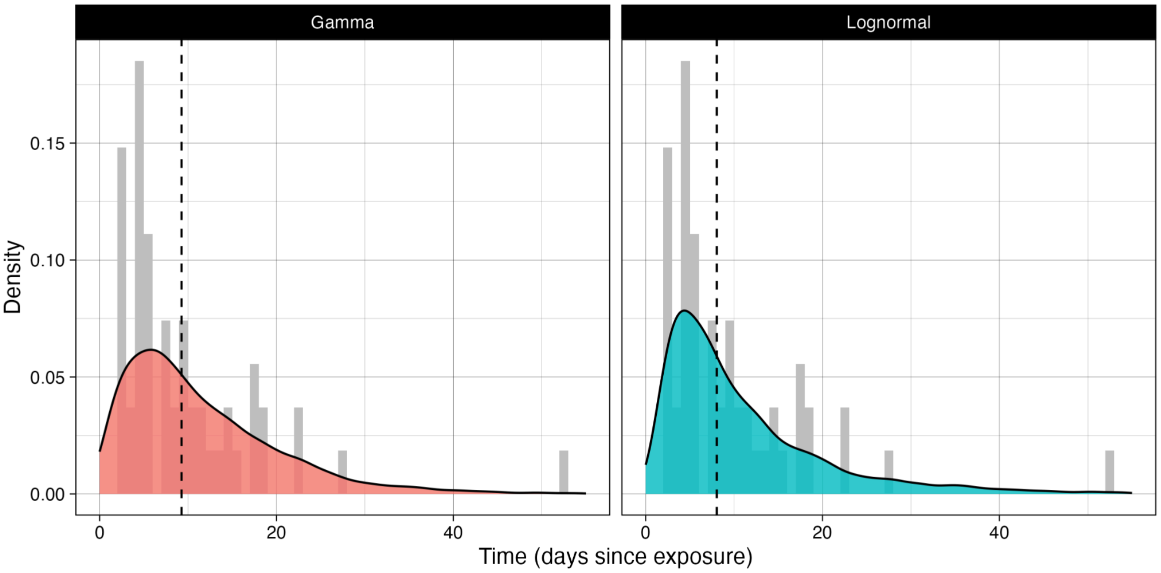
2.2 Serial interval estimation

To estimate serial intervals where appropriate, we used the model presented in Ward et al. 2024 [5] to fit both lognormal and gamma distributions to the number of onsets for a given day from outbreak data reporting the interval between the onset of illness in successive cases. To account for the double censoring we used the R package {primarycensored} [6,7]. Leave-one-out (LOO) analysis was used to select the final parametric model .

2.2.1 LASV

Only one serial interval estimate for LASV was identified during the review stage. With a mean of 7.8 days (SD of 10.7) [8], this was deemed to be biased towards a shorter end of potential serial intervals for Lassa fever, considering the accepted incubation period of 7-21 days [9], which would in turn produce high presymptomatic transmission percentage estimates. We combined the dataset used for the Zhao et al. 2020 estimate (Lo Iacono et al. 2015 [10]) with data collected on Lassa outbreaks as described in **Supplementary Figure S1.** We estimated the serial interval distribution to have a mean of 11.8 days (95% CrI: 1.4–43) when fitted with a lognormal distribution, and a mean of 11.5 days (95% CrI: 0.9–34.6) when fitted with a gamma distribution (**Supplementary Figure S2**).

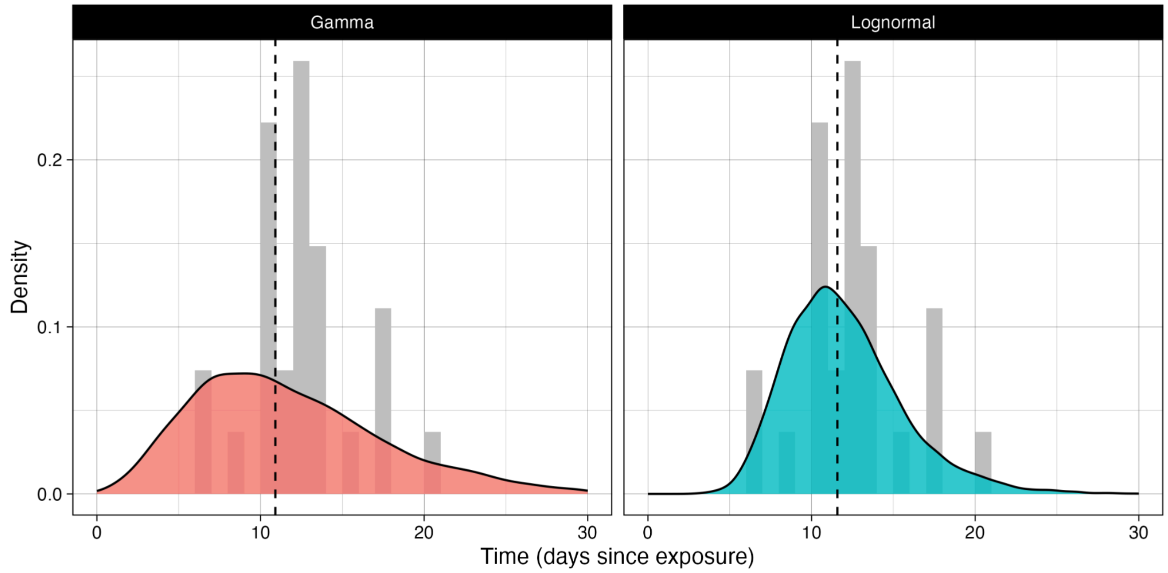
The leave-one-out information criterion (LOOIC) is computed and used to assess the model goodness-of-fit (with smaller values indicating a better fit). In this analysis, the lognormal distribution is the preferred choice as it demonstrates superior predictive accuracy with a smaller LOOIC as compared to the gamma distribution (534.9 vs 557.6).



Supplementary Figure S2. LASV serial interval distribution.
We estimate the serial interval distribution for LASV to have a mean of 11.8 days (95% CrI: 1.4–43) when fitted with a lognormal distribution, and a mean of 11.5 days (95% CrI: 0.9–34.6) when fitted with a gamma distribution

2.2.2 CCHFV

We did not identify any published estimates for CCHFV during the literature search. We collected data on CCHFV outbreaks as described in **Supplementary Figure S1** and estimated the serial interval distribution to have a mean of 12.0 days (95% CrI: 6.5–20.5) when fitted with a lognormal distribution, and a mean of 12.0 days (95% CrI: 3.0–27.2) when fitted with a gamma distribution (**Supplementary Figure S3**). In this analysis, the gamma distribution is the preferred choice as it demonstrates superior predictive accuracy with a smaller LOOIC as compared to the lognormal distribution (135.4 vs 304.1).

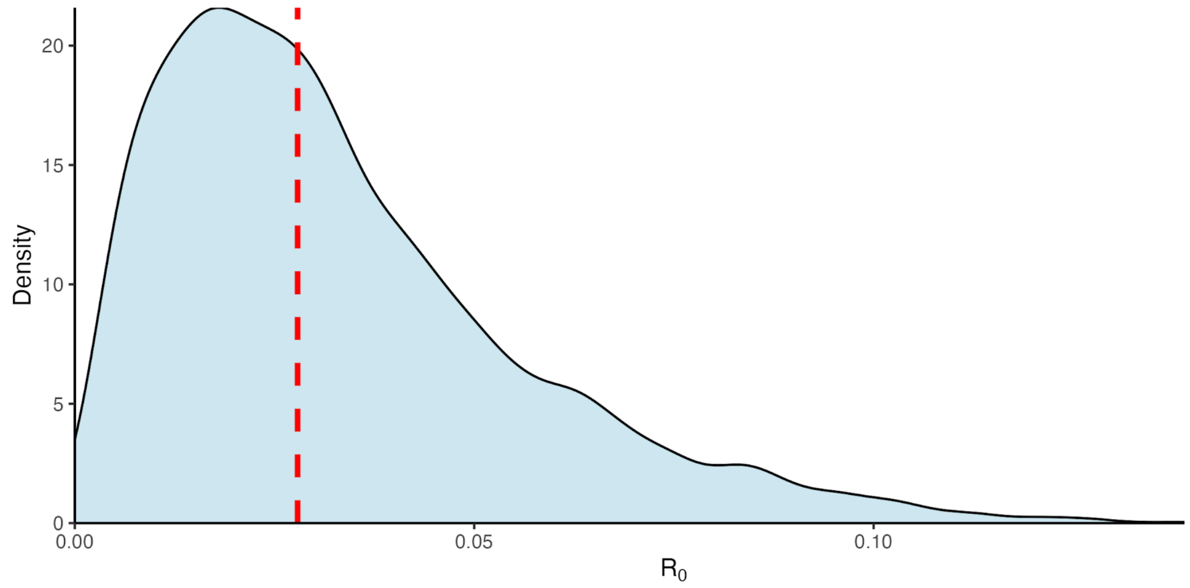


Supplementary Figure S3. CCHFV serial interval distribution.
We estimate the serial interval distribution to have a mean of 12.0 days (95% CrI: 6.5–20.5) when fitted with a lognormal distribution, and a mean of 12.0 days (95% CrI: 3.0–27.2) when fitted with a gamma distribution

2.3 Reproduction number estimation

2.3.1 CCHFV

No basic reproduction number estimates for CCHFV were identified in the literature search. To provide an estimate for the clustering analysis we estimated the basic reproduction number using the package {epichains} [11,12]. The parameters were estimated using Markov chain Monte Carlo (MCMC), implemented in the {MCMCpack package} [13]. We used this to estimate the reproduction number under the assumption of a negative binomial offspring distribution [11]. We used data on CCHFV outbreak clusters to estimate R_0. This data was collected from cases of CCHFV infected in the European union/European Economic Area from 2013–2024 as reported by European Centre for Disease Prevention and Control (ECDC) [14]**.** We assumed cases with unknown transmission were tick-borne cases. We estimated the R_0 for CCHFV with a median of 0.03 (95% CrI: 0.004–0.09) (**Supplementary figure S4)**



Supplementary figure S4. Estimated CCHFV R_0 distribution in the EU from 2013-2024.

Estimated R_0 distribution for CCHFV, red dotted line indicates the median value of 0.03 (95% CrI: 0.004–0.09).

2.4 Percentage of presymptomatic transmission

We estimated the percentage of presymptomatic transmission using published or derived estimates of the incubation period and serial interval for 18 pathogens. The presence of presymptomatic transmission was inferred under the assumption that if the mean serial interval was shorter than the mean incubation period, a portion of transmission likely occurred prior to symptom onset [15].

To quantify this, we plotted the cumulative distribution functions (CDFs) of the incubation period and serial interval distributions for each pathogen. The percentage of presymptomatic transmission was estimated by integrating the difference between the cumulative probabilities of the two distributions:

$$Presymptomatic\ transmission\ (\%) = [F_{serial}(t) - F_{incubation}(t)] \times 100$$

- $F_{serial}(t)$ *is the CDF of the serial interval, representing the probability that transmission has occurred by time t.*
- $F_{incubation}(t)$ *is the CDF of the incubation period, representing the probability that symptom onset has occurred by time t.*

The difference between these cumulative probabilities across time reflects the proportion of transmission events that occurred prior to the onset of symptoms.

3. Clustering

Epidemiological parameters used for clustering included the basic reproduction number (R_0), serial interval, case fatality risk (CFR), and the proportion of presymptomatic transmission. R_0, serial interval, and CFR were summarised using central tendency measures identified during the review stage (see value justification). Serial intervals were recorded in days, while CFR and presymptomatic transmission proportions were expressed as percentages. Transmission modes were recorded as categorical variables. All values and justifications used in clustering are documented in the clustering dataset.

3.1 K-means clustering

For K-means clustering, categorical variables were transformed using one-hot encoding, converting them into binary indicators (0/1). Continuous variables were normalised to ensure comparability, standardizing them to a mean of 0 and a standard deviation of 1.

To determine the optimal number of clusters, we applied the elbow method, a heuristic technique in unsupervised learning. This method identifies the point at which the Within-Cluster Sum of Squares (WCSS) begins to plateau, indicating diminishing returns in additional clusters. The optimal range was estimated to be 4–6 clusters (**Figure 1b)**.

We performed clustering with K = 4, 5, 6 to compare cluster cohesion and separability. The K-means algorithm was applied iteratively, assigning each data point to the nearest cluster centroid and updating centroids until convergence to minimize WCSS. To account for sensitivity to initial centroid placement, the algorithm was run 25 times with different starting points , selecting the solution with the lowest WCSS.

The final solution, K = 5, was selected based on the elbow plot and practical interpretability of the resulting clusters. Cluster centroids were extracted to summarize the mean values of key epidemiological parameters for each cluster.

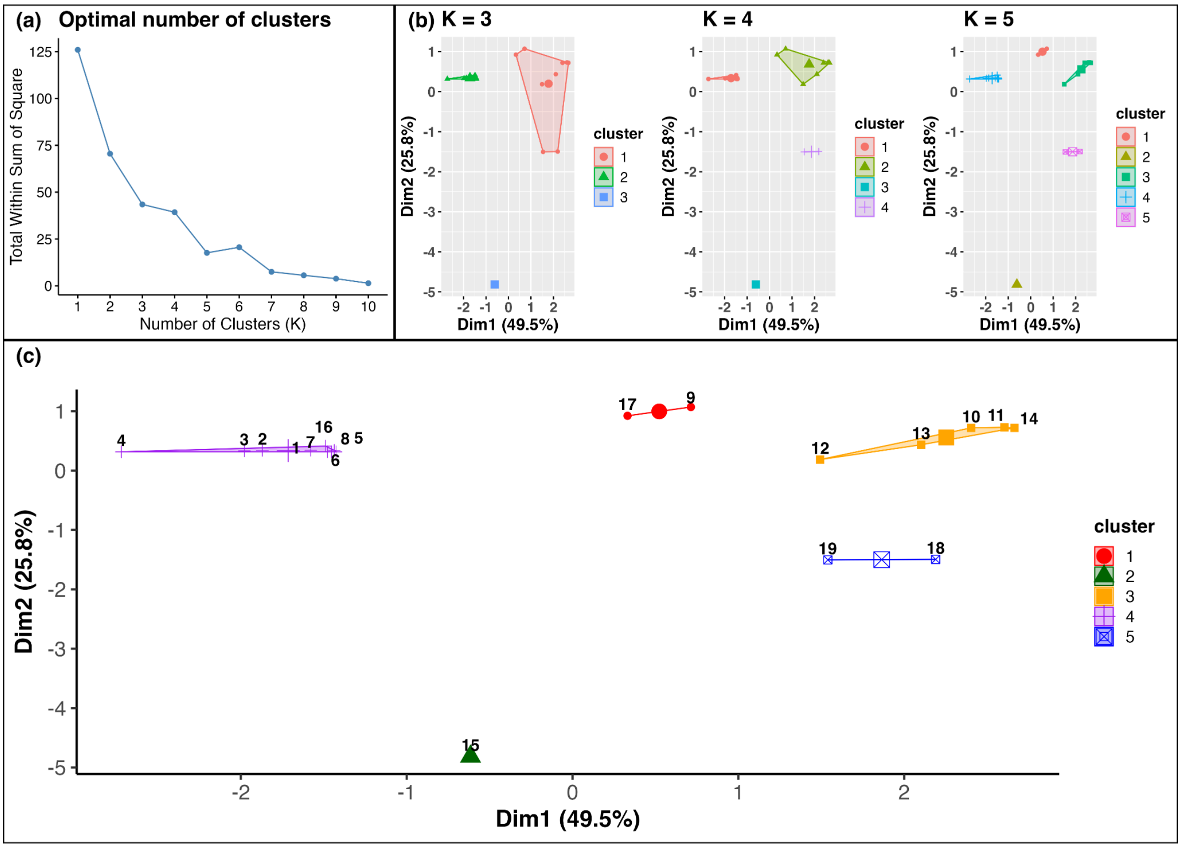
3.2 Hierarchical clustering

For hierarchical clustering, categorical variables were converted into factors to facilitate proper handling. The Gower distance metric was used to measure similarity between pathogens, allowing for the integration of both numerical and categorical data.

Clustering was performed using Ward’s linkage method, which optimises cluster compactness by minimizing the total variance within clusters. Similar to K-means, the elbow method was applied to estimate the optimal number of clusters, which was determined to be between 3 and 4 (**Figure 2b**).

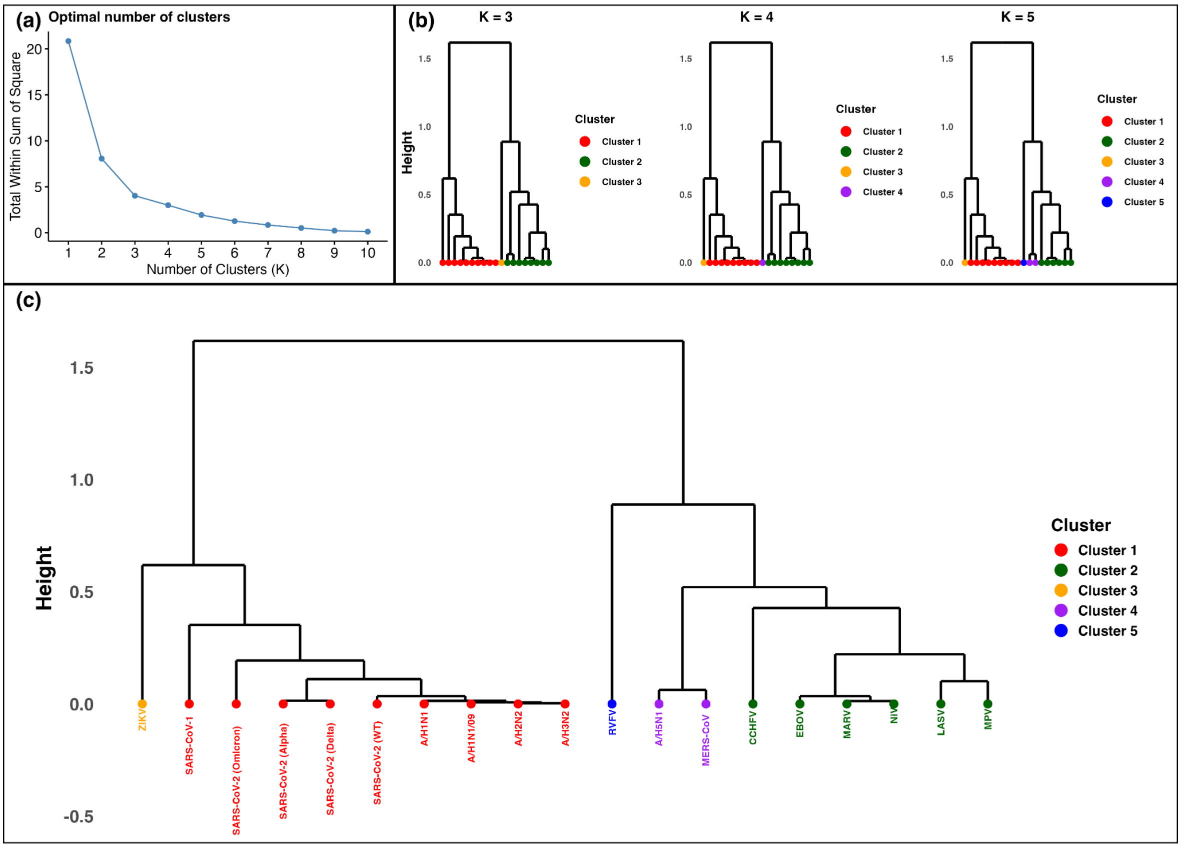
To assess clustering robustness, we explored solutions with K = 3, 4, 5, ultimately selecting K = 4 as the final solution. The mean values of key epidemiological parameters were extracted for each cluster to facilitate interpretation and comparison.

3.3 Clustering sensitivity analysis

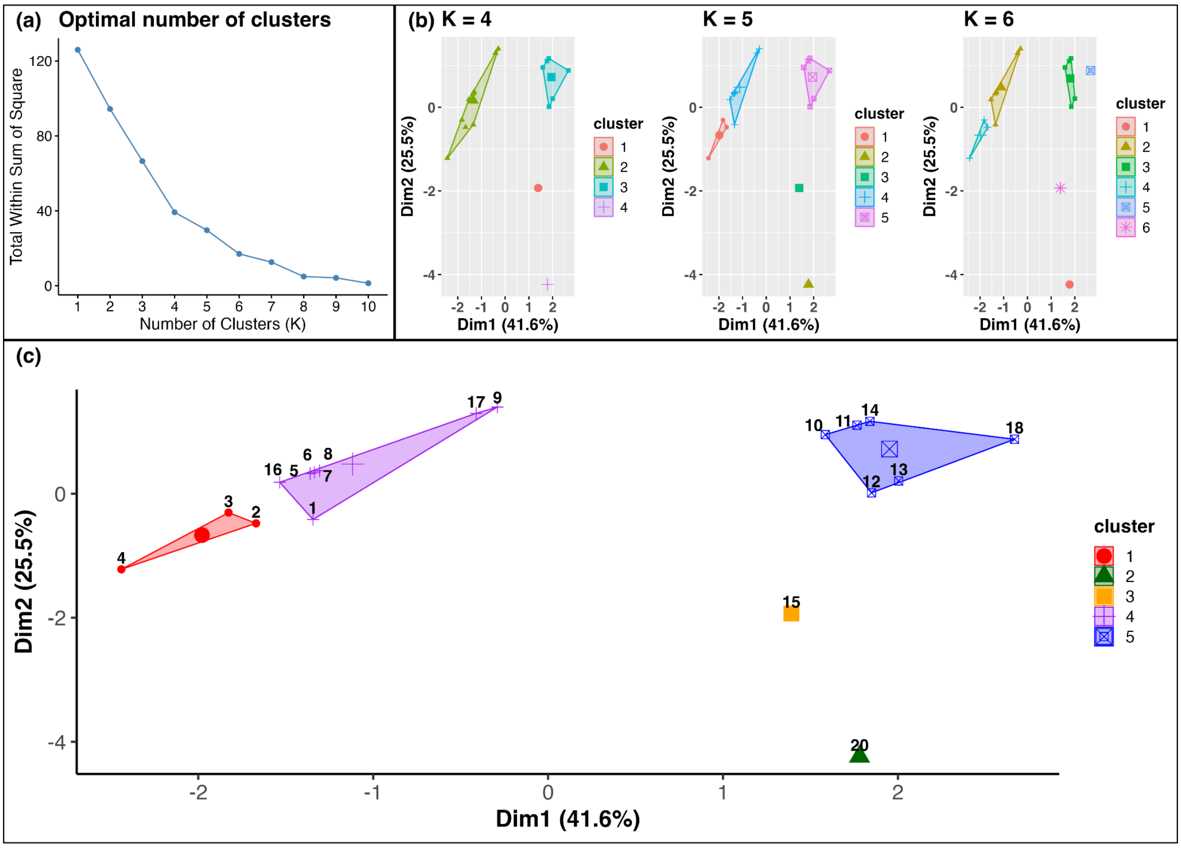


Supplementary figure S4. K-means clustering with R0, CFR and transmission route. (a) Elbow plot used to determine the optimal number of clusters. (b) Clustering analysis using three K values (K = 3,4,5). (c) Final solution retaining K = 5 clusters.

1) SARS-CoV-2 (WT). 2) SARS-CoV-2 (Alpha). 3) SARS-CoV-2 (Delta). 4) SARS-CoV-2 (Omicron). 5) A/H5N1. 6) A/H2N2. 7) A/H3N2. 8) A/H1N1pdm09. 9) A/H5N1. 10) EBOV. 11) MARV. 12) MPV. 13) LASV. 14) NiV. 15) ZIKV. 16) SARS-CoV-1. 17) MERS-CoV. 18) CCHFV. 19) RVFV. Larger unlabeled central points represent cluster centroids. The ZIKV cluster is a cluster of one therefore there is no central point.



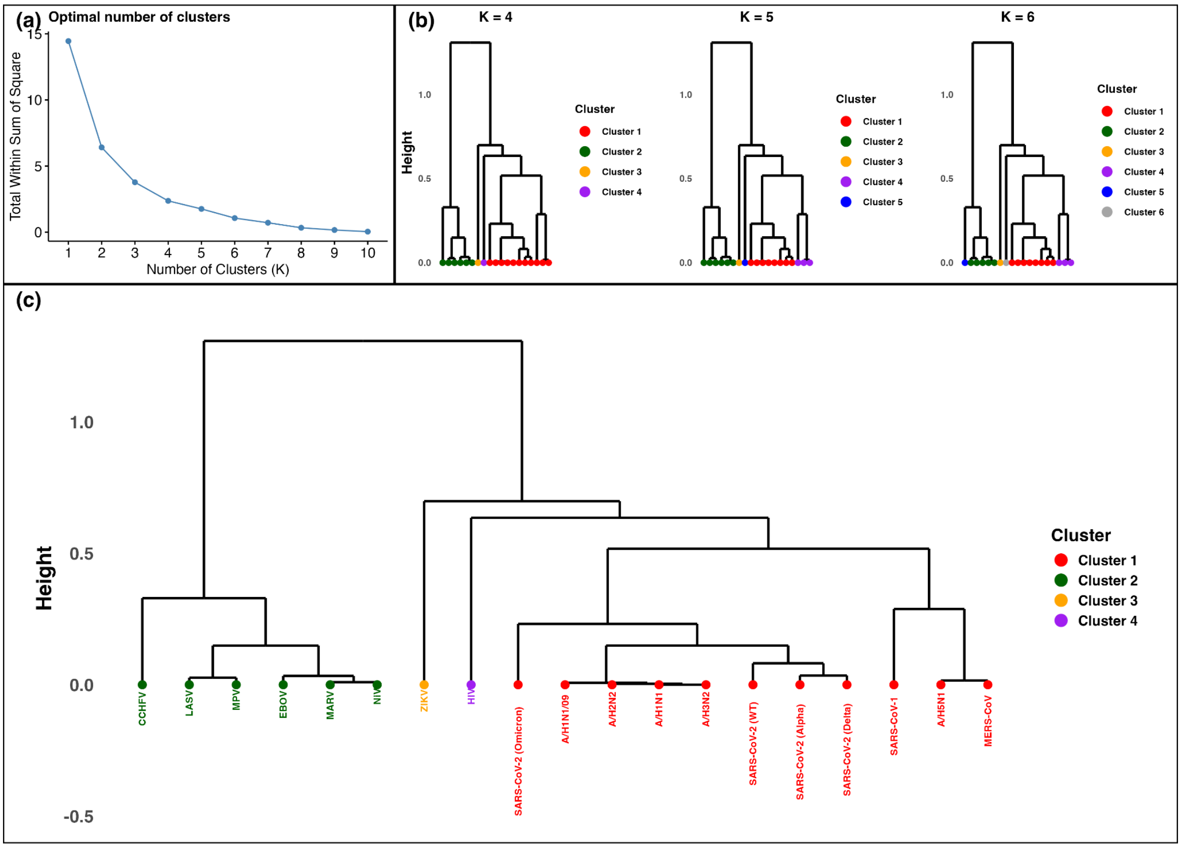
Supplementary figure S5. Hierarchical clustering with R0, CFR and transmission route.
(a) Elbow plot used to determine the optimal number of clusters. (b) Clustering analysis using three K values (K = 3,4,5). (c) Final solution retaining K = 5 clusters.



Supplementary figure S6. K-means clustering including HIV with R0, presymptomatic transmission and transmission route

(a) Elbow plot used to determine the optimal number of clusters. (b) Clustering analysis using three K values (K = 4,5,6). (c) Final solution retaining K = 4 clusters.

1) SARS-CoV-2 (WT). 2) SARS-CoV-2 (Alpha). 3) SARS-CoV-2 (Delta). 4) SARS-CoV-2 (Omicron). 5) A/H5N1. 6) A/H2N2. 7) A/H3N2. 8) A/H1N1pdm09. 9) A/H5N1. 10) EBOV. 11) MARV. 12) MPV. 13) LASV. 14) NiV. 15) ZIKV. 16) SARS-CoV-1. 17) MERS-CoV. 18) CCHFV. 20) HIV. Larger unlabeled central points represent cluster centroids. The ZIKV/HIV cluster is a cluster of one therefore there is no central point.



Supplementary figure S7. Hierarchical clustering including HIV with R0, presymptomatic transmission and transmission route
(a) Elbow plot used to determine the optimal number of clusters. (b) Clustering analysis using three K values (K = 4,5,6). (c) Final solution retaining K = 4 clusters.

4. Computational details

All analyses were run using R [16]. All code to reproduce this report is open on GitHub at https://github.com/oswaldogressani/Blueprint for incubation period estimates and https://github.com/Jward2847/archetypes# for serial interval, reproduction number and percentage of presymptomatic transmission estimates in addition to the clustering analysis.

5. References

[1] Connected Papers n.d. https://www.connectedpapers.com/ (accessed July 9, 2024).
[2] Gressani O. EpiLPS: A Fast and Flexible Bayesian Tool for Estimating Epidemiological Parameters. 2021.
[3] Gressani O, Torneri A, Hens N, Faes C. Flexible Bayesian estimation of incubation times. Am J Epidemiol 2025;194:490–501.
[4] Gressani O, Wallinga J, Althaus CL, Hens N, Faes C. EpiLPS: A fast and flexible Bayesian tool for estimation of the time-varying reproduction number. PLoS Comput Biol 2022;18:e1010618.
[5] Ward J, Lambert JW, Russell TW, Azam JM, Kucharski AJ, Funk S, et al. Estimates of

epidemiological parameters for H5N1 influenza in humans: a rapid review. medRxiv 2024:2024.12.11.24318702. https://doi.org/10.1101/2024.12.11.24318702.

[6] Charniga K, Park SW, Akhmetzhanov AR, Cori A, Dushoff J, Funk S, et al. Best practices for estimating and reporting epidemiological delay distributions of infectious diseases. PLoS Comput Biol 2024;20:e1012520.

[7] Abbott S, Brand S. primarycensored: Primary Event Censored Distributions 2024. https://doi.org/10.5281/zenodo.13632839.

[8] Zhao S, Musa SS, Fu H, He D, Qin J. Large-scale Lassa fever outbreaks in Nigeria: quantifying the association between disease reproduction number and local rainfall. Epidemiol Infect 2020;148:e4.

[9] CDC. About. Lassa Fever 2024. https://www.cdc.gov/lassa-fever/about/index.html (accessed June 26, 2024).

[10] Lo Iacono G, Cunningham AA, Fichet-Calvet E, Garry RF, Grant DS, Khan SH, et al. Using modelling to disentangle the relative contributions of zoonotic and anthroponotic transmission: the case of lassa fever. PLoS Negl Trop Dis 2015;9:e3398.

[11] Azam JM, Funk S, Finger F. epichains: Simulating and Analysing Transmission Chain Statistics Using #> Branching Process Models 2024.

[12] Valle A. howto: How-To Guides For Outbreak Analytics R Packages. 2023.

[13] Martin AD, Quinn KM, Park JH. MCMCpack: Markov Chain Monte Carlo in R. Journal of Statistical Software 2011;42:22. https://doi.org/10.18637/jss.v042.i09.

[14] Cases of Crimean–Congo haemorrhagic fever infected in the EU/EEA, 2013–present. European Centre for Disease Prevention and Control 2021. https://www.ecdc.europa.eu/en/crimean-congo-haemorrhagic-fever/surveillance/cases-eu-since-2013 (accessed February 4, 2025).

[15] He X, Lau EHY, Wu P, Deng X, Wang J, Hao X, et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. Nat Med 2020;26:672–5.

[16] R Core Team. R: A Language and Environment for Statistical Computing 2021.