Supplementary Information

Preparing for the next pandemic takes more than better models: expert consensus recommendations for infectious disease modelling

Leonard Stellbrink et al.

Contents

Supplementary Notes

Supplementary Note 1: Publication characteristics

Supplementary Table 1 describes the 135 included publications by period, type, and study design. Almost half appeared from 2022 onwards, and reviews and systematic reviews together form the largest share, which follows from an eligibility rule that rewards reflection on modelling practice over the reporting of a single model.

Supplementary Table 1: **Publication characteristics of the review corpus.** Characteristics of the 135 included publications. Publication types follow a controlled vocabulary applied to every record, and types with no publications are omitted. Percentages are of the 135 publications and may not sum to 100 due to rounding.

Characteristic	n	%
Publication period		
Before 2010	6	4.4
2010 to 2014	14	10.4
2015 to 2019	25	18.5
2020 to 2021	26	19.3
2022 onward	64	47.4
Publication type		
Research article	11	8.1
Review article	57	42.2
Systematic review	21	15.6
Perspective/Opinion	31	23.0
Editorial	3	2.2
Commentary	6	4.4
Technical report	1	0.7
Policy paper	1	0.7
Other	4	3.0
Primary disease focus		
COVID-19	39	28.9
Influenza	11	8.1
HIV/AIDS	8	5.9

continued on next page

Supplementary Table 1: *(continued)*

Characteristic	n	%
Ebola	3	2.2
Multiple diseases	3	2.2
Non-specific/general	63	46.7
Other	8	5.9
Relevance band		
High (7 to 8)	92	68.1
Moderate (5 to 6)	40	29.6
Low (below 5)	3	2.2

Supplementary Note 2: The included publications

Supplementary Table 2 lists all 135 publications, with the publication type recorded during screening and the themes the term-based codebook (Supplementary Note 18) assigns to each. Six publications match no theme term, which reflects the terseness of the extracted text rather than an absence of relevant content, and is one reason we report the theme shares as lower bounds.

Supplementary Table 2: **Publications included in the systematic review.** The 135 publications included, with the publication type recorded during screening and the themes the term-based codebook (Supplementary Note 18) assigns from the extracted challenge and recommendation text. Themes are not mutually exclusive. A dash means the codebook matched no theme term, not that the publication addresses none.

First author	Year	Title	Type	Themes
Adams et al.	2023	Undoing elimination: Modelling Australia's way out of the COVID-19 pandemic.	Research article	Capacity, Transparency, Uncertainty
Adib et al.	2021	A participatory modelling approach for investigating the spread of COVID-19 in countries of the Eastern Mediterranean Region to support public health decision-making.	Technical report	Capacity, Data, Methods, Uncertainty
Aguas et al.	2020	Modelling the COVID-19 pandemic in context: an international participatory approach.	Perspective/ Opinion	Capacity, Data, Sci-policy, Transparency, Uncertainty

continued on next page

Supplementary Table 2: *(continued)*

First author	Year	Title	Type	Themes
Akman et al.	2021	The Hard Lessons and Shifting Modeling Trends of COVID-19 Dynamics: Multiresolution Modeling Approach.	Perspective/ Opinion	Data, Public comm., Sci-policy, Uncertainty
Al-Tawfiq et al.	2021	Middle East respiratory syndrome coronavirus - The need for global proactive surveillance, sequencing and modeling	Editorial	Capacity
Alahmadi et al.	2020	Influencing public health policy with data-informed mathematical models of infectious diseases: Recent developments and new challenges.	Review article	Capacity, Methods, Public comm., Sci-policy, Transparency, Uncertainty
Alexander et al.	2012	Modeling of wildlife-associated zoonoses: applications and caveats.	Review article	Data, Methods
Ali et al.	2024	Incorporating Social Determinants of Health in Infectious Disease Models: A Systematic Review of Guidelines.	Systematic review	Capacity, Methods, Public comm., Sci-policy
Asplin et al.	2024	Symptom propagation in respiratory pathogens of public health concern: a review of the evidence	Systematic review	–
Avramov et al.	2024	A conceptual health state diagram for modelling the transmission of a (re)emerging infectious respiratory disease in a human population.	Commentary	Data, Uncertainty
Aylett-Bullock et al.	2022	Epidemiological modelling in refugee and internally displaced people settlements: challenges and ways forward.	Review article	Capacity, Data, Methods, Public comm., Reporting, Uncertainty
Barbrook-Johnson et al.	2017	Uses of Agent-Based Modeling for Health Communication: the TELL ME Case Study.	Research article	Data
Bartl	2024	Social and Ethical Implications of Digital Crisis Technologies: Case Study of Pandemic Simulation Models During the COVID-19 Pandemic.	Perspective/ Opinion	–
Basáñez et al.	2012	A research agenda for helminth diseases of humans: modelling for control and elimination.	Review article	Methods, Public comm., Sci-policy

continued on next page

Supplementary Table 2: *(continued)*

First author	Year	Title	Type	Themes
Beauchemin et al.	2011	A review of mathematical models of influenza A infections within a host or cell culture: lessons learned and challenges ahead	Review article	Methods
Becker et al.	2021	Development and dissemination of infectious disease dynamic transmission models during the COVID-19 pandemic: what can we learn from other pathogens and how can we move forward?	Review article	Methods, Public comm., Sci-policy, Transparency, Uncertainty
Bedson et al.	2021	A review and agenda for integrated disease models including social and behavioural factors.	Review article	Capacity, Data, Methods, Sci-policy
Bershteyn et al.	2022	Real-Time Infectious Disease Modeling to Inform Emergency Public Health Decision Making.	Review article	Capacity, Data, Methods, Sci-policy, Uncertainty
Borchering et al.	2024	Responding to the Return of Influenza in the United States by Applying Centers for Disease Control and Prevention Surveillance, Analysis, and Modeling to Inform Understanding of Seasonal Influenza	Perspective/ Opinion	Methods
Borlase et al.	2023	Modelling morbidity for neglected tropical diseases: the long and winding road from cumulative exposure to long-term pathology.	Perspective/ Opinion	Methods, Uncertainty
Bozzani et al.	2021	Building resource constraints and feasibility considerations in mathematical models for infectious disease: A systematic literature review.	Systematic review	Data, Methods, Transparency
Brandt	1989	Policy implications of modelling of the AIDS epidemic.	Perspective/ Opinion	Methods, Uncertainty
Brock et al.	2016	Plasmodium knowlesi transmission: integrating quantitative approaches from epidemiology and ecology to understand malaria as a zoonosis.	Review article	Data, Methods
Caldwell et al.	2021	Vaccines and variants: Modelling insights into emerging issues in COVID-19 epidemiology.	Review article	Uncertainty

continued on next page

Supplementary Table 2: *(continued)*

First author	Year	Title	Type	Themes
Cauchemez et al.	2019	How Modelling Can Enhance the Analysis of Imperfect Epidemic Data.	Review article	Capacity, Data, Public comm.
Charniga et al.	2024	Nowcasting and forecasting the 2022 U.S. mpox outbreak: Support for public health decision making and lessons learned.	Research article	Methods, Public comm., Uncertainty
Childs et al.	2015	Modelling challenges in context: lessons from malaria, HIV, and tuberculosis.	Review article	Data
Chowell et al.	2017	Perspectives on model forecasts of the 2014-2015 Ebola epidemic in West Africa: lessons and the way forward.	Perspective/ Opinion	Data, Methods
Chowell et al.	2024	Investigating and forecasting infectious disease dynamics using epidemiological and molecular surveillance data.	Review article	Data, Methods, Uncertainty
Comito et al.	2022	Artificial intelligence for forecasting and diagnosing COVID-19 pandemic: A focused review.	Systematic review	Data, Methods, Sci-policy
Cooper et al.	2021	How Good is the Science That Informs Government Policy? A Lesson From the U.K.'s Response to 2020 CoV-2 Outbreak.	Perspective/ Opinion	Methods, Public comm., Sci-policy, Uncertainty
Cori et al.	2024	Inference of epidemic dynamics in the COVID-19 era and beyond.	Review article	Data, Methods, Transparency
Cuadros et al.	2024	Advancing Public Health Surveillance: Integrating Modeling and GIS in the Wastewater-Based Epidemiology of Viruses, a Narrative Review	Review article	Data, Methods, Public comm.
Dangerfield et al.	2022	Challenges of integrating economics into epidemiological analysis of and policy responses to emerging infectious diseases.	Perspective/ Opinion	Sci-policy, Uncertainty
Dangerfield et al.	2023	Getting the most out of maths: How to coordinate mathematical modelling research to support a pandemic, lessons learnt from three initiatives that were part of the COVID-19 response in the UK.	Perspective/ Opinion	Capacity, Methods

continued on next page

Supplementary Table 2: *(continued)*

First author	Year	Title	Type	Themes
Darquenne et al.	2022	Aerosol Transport Modeling: The Key Link Between Lung Infections of Individuals and Populations	Review article	Uncertainty
Dattner et al.	2022	The role of statisticians in the response to COVID-19 in Israel: a holistic point of view	Perspective/ Opinion	Capacity, Data, Public comm., Sci-policy, Uncertainty
De Angelis et al.	2015	Four key challenges in infectious disease modelling using data from multiple sources.	Perspective/ Opinion	Data, Methods
Delva et al.	2012	HIV treatment as prevention: principles of good HIV epidemiology modelling for public health decision-making in all modes of prevention and evaluation.	Review article	Capacity, Methods, Transparency, Uncertainty
Dembek et al.	2018	Best practice assessment of disease modelling for infectious disease outbreaks.	Review article	Data, Methods, Sci-policy, Uncertainty
Deng et al.	2022	Mathematical Models Supporting Control of COVID-19	Review article	Data, Methods
Desai et al.	2019	Real-time Epidemic Forecasting: Challenges and Opportunities.	Review article	Capacity, Data, Methods, Public comm., Uncertainty
Doms et al.	2018	Assessing the Use of Influenza Forecasts and Epidemiological Modeling in Public Health Decision Making in the United States.	Research article	–
Dorjee et al.	2013	A Review of Simulation Modelling Approaches Used for the Spread of Zoonotic Influenza Viruses in Animal and Human Populations	Systematic review	Methods
Drake et al.	2015	Transmission Models of Historical Ebola Outbreaks.	Review article	Methods
El Morr et al.	2024	AI-based epidemic and pandemic early warning systems: A systematic scoping review.	Systematic review	Capacity, Data, Public comm.
Erzen et al.	2020	Key challenges in modelling an epidemic - What have we learned from the COVID-19 epidemic so far.	Editorial	Methods, Public comm., Uncertainty

continued on next page

Supplementary Table 2: *(continued)*

First author	Year	Title	Type	Themes
Espinosa et al.	2024a	The value of mathematical modelling approaches in epidemiology for public health decision making; [La utilidad de los modelos matemáticos en epidemiología para la toma de decisiones en salud pública]	Perspective/ Opinion	Data, Public comm., Sci-policy, Uncertainty
Espinosa et al.	2024b	Predictive models for health outcomes due to SARS-CoV-2, including the effect of vaccination: a systematic review.	Systematic review	Methods
Fischer et al.	2016	CDC Grand Rounds: Modeling and Public Health Decision-Making.	Other	Capacity, Data
Fisman	2023	When the role of uncertainty is. . . uncertain	Commentary	Capacity, Data, Uncertainty
Fitzpatrick et al.	2019	Modelling microbial infection to address global health challenges.	Perspective/ Opinion	Data, Methods, Public comm., Uncertainty
Gandon et al.	2016	Forecasting Epidemiological and Evolutionary Dynamics of Infectious Diseases.	Review article	Data, Methods, Uncertainty
Gawande et al.	2025	The role of artificial intelligence in pandemic responses: from epidemiological modeling to vaccine development.	Review article	Data
Giddings et al.	2023	Infectious Disease Modelling of HIV Prevention Interventions: A Systematic Review and Narrative Synthesis of Compartmental Models.	Systematic review	Methods, Transparency, Uncertainty
Glennon et al.	2021	Challenges in modeling the emergence of novel pathogens.	Review article	Data, Methods, Sci-policy, Transparency, Uncertainty
Grad et al.	2012	Cholera modeling: challenges to quantitative analysis and predicting the impact of interventions.	Perspective/ Opinion	Methods, Uncertainty
Grassly et al.	2008	Mathematical models of infectious disease transmission.	Review article	Methods
Hadley et al.	2021	Challenges on the interaction of models and policy for pandemic control.	Perspective/ Opinion	Capacity, Methods, Reporting, Transparency, Uncertainty

continued on next page

Supplementary Table 2: *(continued)*

First author	Year	Title	Type	Themes
Hamilton et al.	2024	Research gaps and priorities for quantitative microbial risk assessment (QMRA)	Research article	Capacity, Methods, Public comm., Sci-policy, Transparency, Uncertainty
Heesterbeek et al.	2015	Modeling infectious disease dynamics in the complex landscape of global health.	Review article	Data, Methods, Sci-policy
Heneghan et al.	2022	Why COVID-19 modelling of progression and prevention fails to translate to the real-world.	Research article	Methods, Reporting, Transparency, Uncertainty
Hillmer et al.	2021	Ontario's COVID-19 Modelling Consensus Table: mobilizing scientific expertise to support pandemic response.	Commentary	–
James et al.	2021	The Use and Misuse of Mathematical Modeling for Infectious Disease Policymaking: Lessons for the COVID-19 Pandemic.	Review article	Methods, Sci-policy, Uncertainty
Jit et al.	2024	Informing Public Health Policies with Models for Disease Burden, Impact Evaluation, and Economic Evaluation.	Review article	Capacity, Methods, Public comm., Reporting, Sci-policy, Transparency, Uncertainty
Kakehashi et al.	2024	Contributions and problems of mathematical models in COVID-19 prevention in Japan	Review article	Data, Uncertainty
Keeling et al.	2009	Mathematical modelling of infectious diseases.	Review article	Methods, Uncertainty
Kimani et al.	2022	Infectious disease modelling for SARS-CoV-2 in Africa to guide policy: A systematic review.	Systematic review	Capacity, Data, Methods, Sci-policy, Uncertainty
Knight et al.	2016	Bridging the gap between evidence and policy for infectious diseases: How models can aid public health decision-making.	Review article	Methods, Public comm., Sci-policy, Transparency, Uncertainty
Kong et al.	2022	Compartmental structures used in modeling COVID-19: a scoping review.	Systematic review	Methods
Kretzschmar	2020	Disease modeling for public health: added value, challenges, and institutional constraints	Perspective/ Opinion	Capacity, Sci-policy, Transparency

continued on next page

Supplementary Table 2: *(continued)*

First author	Year	Title	Type	Themes
Kretzschmar et al.	2008	New challenges for mathematical and statistical modeling of HIV and hepatitis C virus in injecting drug users.	Review article	Methods, Uncertainty
Kretzschmar et al.	2022	Challenges for modelling interventions for future pandemics.	Perspective/ Opinion	Capacity, Data, Methods, Sci-policy, Uncertainty
Lal	2016	Spatial Modelling Tools to Integrate Public Health and Environmental Science, Illustrated with Infectious Cryptosporidiosis.	Review article	Capacity, Data
Le Rutte et al.	2024	A case for ongoing structural support to maximise infectious disease modelling efficiency for future public health emergencies: A modelling perspective.	Perspective/ Opinion	Public comm., Sci-policy
Lee et al.	2013	Modelling during an emergency: the 2009 H1N1 influenza pandemic.	Review article	Data, Methods, Transparency
Lee et al.	2023	Mathematical Modeling of COVID-19 Transmission and Intervention in South Korea: A Review of Literature.	Systematic review	Methods
Li et al.	2024	Mathematical models and analysis tools for risk assessment of unnatural epidemics: a scoping review	Systematic review	Data, Methods
Løchen et al.	2020	Dynamic transmission models and economic evaluations of pneumococcal conjugate vaccines: a quality appraisal and limitations.	Review article	Reporting, Transparency, Uncertainty
Mandal et al.	2022	'Imperfect but useful': pandemic response in the Global South can benefit from greater use of mathematical modelling.	Other	Capacity, Methods, Uncertainty
Maziarz et al.	2020	Agent-based modelling for SARS-CoV-2 epidemic prediction and intervention assessment: A methodological appraisal.	Perspective/ Opinion	Methods, Reporting, Sci-policy, Transparency, Uncertainty
McBryde et al.	2020	Role of modelling in COVID-19 policy development.	Review article	Capacity, Sci-policy

continued on next page

Supplementary Table 2: *(continued)*

First author	Year	Title	Type	Themes
McKee et al.	2025	The power of artificial intelligence for managing pandemics: A primer for public health professionals.	Review article	Capacity, Data, Reporting
McVernon et al.	2007	Model answers or trivial pursuits? The role of mathematical models in influenza pandemic preparedness planning.	Review article	Uncertainty
Meehan et al.	2020	Modelling insights into the COVID-19 pandemic.	Review article	Capacity, Data, Methods, Uncertainty
Metcalf et al.	2017	Opportunities and challenges in modeling emerging infectious diseases.	Review article	Data, Public comm., Uncertainty
Mihaljevic et al.	2024	An inaugural forum on epidemiological modeling for public health stakeholders in Arizona	Research article	Capacity, Data, Public comm., Sci-policy
Moghadas et al.	2011	Canada in the face of the 2009 H1N1 pandemic.	Other	–
Naidoo et al.	2024	Incorporating social vulnerability in infectious disease mathematical modelling: a scoping review.	Systematic review	Capacity, Methods, Public comm., Transparency
Ndeffo-Mbah et al.	2018	Dynamic Models of Infectious Disease Transmission in Prisons and the General Population	Systematic review	Data, Methods, Uncertainty
Nelson et al.	2019	Fogarty International Center collaborative networks in infectious disease modeling: Lessons learnt in research and capacity building.	Review article	Capacity, Sci-policy
Niv-Yagoda et al.	2022	The role of models as a decision-making support tool rather than a guiding light in managing the COVID-19 pandemic.	Research article	Public comm., Sci-policy, Uncertainty
O'Neill	2010	Introduction and snapshot review: Relating infectious disease transmission models to data	Review article	Data, Methods
Olesen et al.	2023	Infectious Disease Modeling: Recommendations for Public Health Decision-Makers	Perspective/ Opinion	Sci-policy, Uncertainty

continued on next page

Supplementary Table 2: *(continued)*

First author	Year	Title	Type	Themes
Opatowski et al.	2018	Influenza interaction with cocirculating pathogens and its impact on surveillance, pathogenesis, and epidemic profile: A key role for mathematical modelling.	Review article	Data, Methods
Pagel et al.	2022	Role of mathematical modelling in future pandemic response policy.	Perspective/ Opinion	Data, Public comm., Sci-policy
Panovska-Griffiths et al.	2022	Technical challenges of modelling real-life epidemics and examples of overcoming these.	Editorial	Data, Public comm., Reporting, Transparency, Uncertainty
Pillai et al.	2024	Agent-based modeling of the COVID-19 pandemic in Florida.	Research article	Data, Methods, Sci-policy
Punyacharoensin et al.	2011	Mathematical models for the study of HIV spread and control amongst men who have sex with men.	Systematic review	Data, Methods, Sci-policy, Transparency, Uncertainty
Quaife et al.	2022	Considering equity in priority setting using transmission models: Recommendations and data needs.	Policy paper	Capacity, Methods, Transparency
Rahman et al.	2021	A Review of COVID-19 Modelling Strategies in Three Countries to Develop a Research Framework for Regional Areas.	Review article	Capacity, Data, Methods, Sci-policy, Uncertainty
Rasheed et al.	2021	COVID-19 in the Age of Artificial Intelligence: A Comprehensive Review.	Review article	Capacity, Data, Transparency
Rennie et al.	2024	Ethics of Mathematical Modeling in Public Health: The Case of Medical Male Circumcision for HIV Prevention in Africa	Perspective/ Opinion	Public comm., Sci-policy, Transparency, Uncertainty
Rhodes et al.	2020	Modelling the pandemic: attuning models to their contexts.	Perspective/ Opinion	Public comm.
Riley et al.	2015	Five challenges for spatial epidemic models.	Perspective/ Opinion	Methods
Rivers et al.	2019	Using "outbreak science" to strengthen the use of models during epidemics.	Commentary	Capacity, Data, Sci-policy, Uncertainty
Rossiter et al.	2024	The role of economic evaluation in modelling public health and social measures for pandemic policy: a systematic review	Systematic review	Reporting, Transparency

continued on next page

Supplementary Table 2: *(continued)*

First author	Year	Title	Type	Themes
Round et al.	2024	Improving Transparency of Decision Models Through the Application of Decision Analytic Models with Omitted Objects Displayed (DAMWOOD)	Research article	Methods, Public comm., Reporting, Transparency
Ryan et al.	2023	The current landscape of software tools for the climate-sensitive infectious disease modelling community	Systematic review	Capacity, Data, Reporting
Sarmiento Varón et al.	2023	The role of machine learning in health policies during the COVID-19 pandemic and in long COVID management.	Review article	Capacity
Shah et al.	2020	Unfolding trends of COVID-19 transmission in India: Critical review of available Mathematical models	Systematic review	Uncertainty
Shanmugam et al.	2024	Comprehending symmetry in epidemiology: A review of analytical methods and insights from models of COVID-19, Ebola, Dengue, and Monkeypox	Review article	Data, Methods, Sci-policy, Uncertainty
Stachel et al.	2021	Modeling transmission of pathogens in healthcare settings.	Review article	Methods, Public comm., Sci-policy
Steinberg et al.	2022	The role of models in the covid-19 pandemic.	Other	Data, Public comm., Sci-policy, Uncertainty
Stockdale et al.	2022	The potential of genomics for infectious disease forecasting.	Perspective/ Opinion	Data, Methods, Reporting, Transparency
Stover	2000	Influence of mathematical modeling of HIV and AIDS on policies and programs in the developing world.	Review article	Sci-policy
Sun et al.	2023	The application of simulation methods during the COVID-19 pandemic: A scoping review.	Systematic review	Capacity, Data, Methods
Swallow et al.	2022	Challenges in estimation, uncertainty quantification and elicitation for pandemic modelling.	Review article	Capacity, Data, Methods, Uncertainty
Teerawattananon et al.	2022	Recalibrating the notion of modelling for policymaking during pandemics.	Perspective/ Opinion	Capacity, Methods, Public comm., Reporting

continued on next page

Supplementary Table 2: *(continued)*

First author	Year	Title	Type	Themes
Thompson et al.	2020	Key questions for modelling COVID-19 exit strategies.	Perspective/ Opinion	Capacity, Data, Methods, Sci-policy, Uncertainty
Thompson et al.	2022	A framework for considering the utility of models when facing tough decisions in public health: a guideline for policy-makers	Commentary	Methods, Sci-policy, Transparency, Uncertainty
Tordoff et al.	2024	Including transgender populations in mathematical models for HIV treatment and prevention: current barriers and policy implications.	Commentary	Data, Methods
Tucker et al.	2023	Agent-based social simulations for health crises response: utilising the everyday digital health perspective.	Perspective/ Opinion	Data, Public comm., Sci-policy
Van Kerkhove et al.	2012	Epidemic and intervention modelling–a scientific rationale for policy decisions? Lessons from the 2009 influenza pandemic.	Perspective/ Opinion	Data, Sci-policy, Uncertainty
Viboud et al.	2018	The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt.	Research article	Methods, Sci-policy
Wang et al.	2022	Simulation and forecasting models of COVID-19 taking into account spatio-temporal dynamic characteristics: A review.	Systematic review	Data, Methods, Sci-policy, Transparency
Watson et al.	2015	A review of typhoid fever transmission dynamic models and economic evaluations of vaccination.	Review article	Transparency
Weltz et al.	2022	Reinforcement Learning Methods in Public Health.	Review article	Public comm.
Woolhouse	2011	How to make predictions about future infectious disease risks.	Review article	Data, Methods, Public comm., Sci-policy, Uncertainty
Wu et al.	2011	The use of mathematical models to inform influenza pandemic preparedness and response.	Review article	Data, Uncertainty

continued on next page

Supplementary Table 2: *(continued)*

First author	Year	Title	Type	Themes
Xia et al.	2024	Canada's provincial COVID-19 pandemic modelling efforts: A review of mathematical models and their impacts on the responses.	Review article	Capacity, Data, Methods, Transparency
Ye et al.	2025	Integrating artificial intelligence with mechanistic epidemiological modeling: a scoping review of opportunities and challenges	Systematic review	Capacity, Methods, Public comm.
Ying et al.	2014	Modeling the implementation of universal coverage for HIV treatment as prevention and its impact on the HIV epidemic.	Review article	–
Zang et al.	2019	Structural Design and Data Requirements for Simulation Modelling in HIV/AIDS: A Narrative Review	Review article	Data, Methods
Zelner et al.	2022	There are no equal opportunity infectors: Epidemiological modelers must rethink our approach to inequality in infection risk.	Perspective/ Opinion	Capacity, Sci-policy

Supplementary Note 3: Theme coverage summary

Supplementary Table 3 gives the share of the 135 included publications that address each of the eight themes. We synthesised the themes from the extracted challenge and recommendation text after the statement set was fixed, so they describe the literature rather than explaining how the statements were generated; Supplementary Note 18 gives the term list and the method. Themes are not mutually exclusive, and a publication counts towards a theme only where the extracted text matches one of that theme's terms, so the shares are lower bounds on what each publication discusses.

Supplementary Note 4: Statement-level consensus outcomes

Supplementary Table 4 lists all 40 statements in full, with the Round 4 consensus outcome for each and the Round 3 comparison. Here and throughout the Supplementary Information, statements appear in the wording the panel

Supplementary Table 3: **Theme coverage across the reviewed literature.** Coverage of the eight themes across the 135 included publications. A publication counts as addressing a theme where the challenge and recommendation text extracted from it matches a term from that theme's list, given in Supplementary Note 18. Themes are not mutually exclusive, so shares sum to more than 100. Shares are lower bounds: the term list recovers 0.62 of the themes hand labelling finds, so coverage is understated rather than inflated.

Theme	**Publications**	**%**
Model development and methodological improvement	81	60
Data governance, sharing, and infrastructure	67	50
Uncertainty quantification and communication	63	47
Science-policy interface	48	36
Capacity building and global equity	44	33
Public and stakeholder communication	36	27
Model transparency and reproducibility	31	23
Reporting standards and documentation	13	10

rated, which keeps US spelling. Four abbreviations appear in the wording of the statements: FAIR (findable, accessible, interoperable, and reusable) and FAIR4RS (FAIR principles for research software) in B02, API (application programming interface) in C07 and C08, and LMIC (low- and middle-income country) in D02. Two features of the rated wording are worth flagging; we altered neither because the panel voted on this text. B04 and B08 share one sentence verbatim, on balancing data privacy protections against emergency data-sharing needs, so that proposition was rated within two statements. The closing sentence of B03, which requires frameworks to be established in advance of emergencies, refers back to the data access agreements and synthetic data mentioned earlier in the same statement, rather than to anything outside it. Of the 40 statements, 20 had already reached consensus at Round 3 under the final numbering, and the remaining 20 reached it at Round 4. Under the Round 3 numbering, before the two merges, 21 of the 42 statements met the criterion. No median fell between the rounds, and no statement moved out of consensus. Dispersion moved the other way in one case: the IQR of B05 rose from 0 at Round 3 to 1 at Round 4, leaving it within the consensus band. The reduction in dispersion was most pronounced for the sampling statement (B07), which was the most divisive at Round 3

(interquartile range (IQR) of 4.0) and reached a clear consensus after rewording. Because Round 4 drew on a broader and partly different panel, we interpret these movements descriptively rather than as paired changes. Extended Data Fig. 1 of the main article shows the per-statement change in dispersion between the two quantitative rounds.

Supplementary Table 4: **Statement-level agreement across the final two rounds.** All 40 statements at both quantitative rounds. The R3 and R4 columns give the median rating with the interquartile range (IQR) in parentheses, and the R3 figures pool the two pairs of statements merged after Round 3. n is the number of panellists who rated each statement, excluding "Not qualified to respond" and non-responses, and varies because partial completions were retained. Consensus is based on the IQR of the six-point ratings: IQR ≤ 1 (consensus), $1 <$ IQR ≤ 1.5 (near-consensus), IQR > 1.5 (non-consensus).

ID	Statement	R3	R4	n (R4)	Status
Section A: Model design and complexity					
A01	The suitability of an epidemiological model depends on the question it addresses, although certain quality criteria, such as mathematical correctness, reproducibility, and transparency, apply regardless of the specific research question.	6 (1)	6 (0)	51	Consensus
A02	There is no single model that fits all scenarios. A limited set of well-developed models can address multiple questions, and existing models can often be adapted or extended in crisis situations rather than requiring an entirely new model for each question, provided adequate data are collected.	5 (2)	6 (1)	51	Consensus

continued on next page

Supplementary Table 4: *(continued)*

ID	Statement	R3	R4	n (R4)	Status
A03	We gain deeper insights by integrating diverse models that differ in structure, assumptions, and methodological techniques. Convergence on similar results across independently constructed models increases robustness and plausibility, while divergences illuminate uncertainty, reveal hidden assumptions, and identify areas requiring further investigation.	6 (0)	6 (0)	50	Consensus
A04	In favorable conditions, simple models can illuminate core mechanisms and inform policymakers about future developments, potential mitigations, and their effectiveness and costs. Their simplifying assumptions and boundary conditions should be clearly communicated to policymakers to avoid misinterpretations, and they should be complemented by more complex analyses when needed.	6 (1)	6 (1)	50	Consensus
A05	During pandemics driven by fast-spreading diseases, speed is essential to provide timely insights that support rapid decision-making. Rapid-response modeling should be accompanied by clear communication of uncertainties and caveats, at minimum qualitatively. Formal uncertainty quantification should be pursued where feasible.	5 (2)	6 (0)	50	Consensus
A06	There is a trade-off between model complexity and completeness, so simplifying assumptions should not omit factors critical to the specific question. Sensitivity analysis should be used to determine which simplifications are acceptable.	6 (1)	6 (1)	50	Consensus

continued on next page

Supplementary Table 4: *(continued)*

ID	Statement	R3	R4	n (R4)	Status
A07	Model development should follow an iterative approach with regular evaluation processes, recognizing that not all factors are known at the outset and that requirements may evolve as data accumulate. The pace of model updates should be adapted to the context and urgency of the situation.	5 (2)	6 (1)	49	Consensus
A08	Managing many parameter combinations in complex models poses significant challenges. When the number of parameters to be estimated exceeds what the available data can support, identifiability becomes a problem, making reliable estimation difficult.	5.5 (1)	6 (1)	49	Consensus
A09	Simplified models that focus on key parameters can clarify specific mechanisms when their scope and limitations are made explicit. Where feasible, their outputs should be compared with empirical data and, when available, with more comprehensive frameworks.	5 (2)	6 (1)	49	Consensus
A10	Integrating socio-economic and behavioral factors into epidemiological models can be essential for capturing real-world transmission dynamics and intervention effects, and their inclusion should be justified based on the question being addressed. Doing so, however, increases data and calibration demands, and reliable behavioral data remain scarce, requiring careful validation to balance explanatory richness with predictive reliability.	5 (2)	5.5 (1)	50	Consensus

continued on next page

Supplementary Table 4: *(continued)*

ID	Statement	R3	R4	n (R4)	Status
A11	Agent-based models offer high-resolution modeling of population heterogeneity and spatial dynamics that are difficult to capture with standard compartmental models. They require more setup, calibration, and computational resources. While they can be used for short-term prediction, their particular strength lies in exploring intervention scenarios and system behavior under policy changes. Hybrid approaches combining compartmental and agent-based elements may also be considered.	5 (1.25)	5 (1)	45	Consensus
A12	To support government advice during a pandemic, models should be informed by the best available data on the epidemic state and virus variants to predict policy-relevant outcomes including case numbers, hospitalizations, and deaths. They should incorporate behavioral factors where they influence transmission and consider economic and psychological effects when they meaningfully influence policy decisions. The scope of models should be matched to the decision context and available data.	5 (1)	6 (1)	49	Consensus
A13	Calibration reporting should be explicit, including which calibration method was used, which parameters were estimated, data sources, assumptions, and the associated uncertainties, to support reproducibility and understanding of model performance.	6 (0)	6 (0)	49	Consensus
Section B: Data availability and quality					

continued on next page

Supplementary Table 4: *(continued)*

ID	Statement	R3	R4	n (R4)	Status
B01	Timely data availability is important for effective modeling and decision support. When quality trade-offs are necessary, document how data limitations influence model outputs and clearly communicate uncertainties to decision-makers. In rapid decision settings, preliminary data or best estimates may be used, but analyses should explicitly state the uncertainties and caveats.	6 (0)	6 (0)	49	Consensus
B02	Open sharing of models, data, and code, along with uncertainties and assumptions, in line with the FAIR (and FAIR4RS) principles, both within the scientific community and with policymakers, improves trust, reproducibility, and the practical uptake of model results.	6 (1)	6 (1)	49	Consensus
B03	Open sharing has limitations, including privacy concerns, particularly for individual-level data, which can be mitigated by data access agreements or synthetic data. Aggregate-level data (e.g., hospital admissions, bed occupancy) generally poses minimal privacy risk and should be shared openly by default. These frameworks should be established in advance of emergencies.	4 (1)	6 (1)	49	Consensus
B04	Socio-behavioral data collection should be integrated alongside clinical surveillance from the outset, with the scope adapted to the pathogen and transmission context. Data privacy protections should balance emergency data-sharing needs.	5 (2)	6 (1)	48	Consensus

continued on next page

Supplementary Table 4: *(continued)*

ID	Statement	R3	R4	n (R4)	Status
B05	Non-traditional data sources, such as digital platforms, mobility tracking, search queries, and participatory surveillance, can complement conventional surveillance and should be considered from the outset, with appropriate privacy safeguards.	6 (0)	6 (1)	48	Consensus
B06	Sentinel surveillance and population-representative surveys have complementary goals and both are needed. The choice between them should depend on the specific modeling question.	6 (1)	6 (1)	47	Consensus
B07	Multiple sampling approaches are needed for comprehensive surveillance. Community-based sampling can provide timely prevalence data, while hospital-based sampling, though biased for population prevalence estimation, is valuable for understanding disease severity and health system burden. Wastewater surveillance can provide privacy-preserving, complementary data. All data sources carry inherent biases that should be transparently characterized and accounted for in analysis.	3.5 (4)	6 (1)	48	Consensus
B08	Ethical frameworks should explicitly address inequity in data collection and access to benefits. Data privacy protections should balance emergency data-sharing needs, and predefined ethical safeguards should be established prior to a pandemic.	6 (2)	6 (0)	46	Consensus
Section C: Uncertainty, validation, and communication					

continued on next page

Supplementary Table 4: *(continued)*

ID	Statement	R3	R4	n (R4)	Status
C01	It is important to provide context for model outputs, including estimates of uncertainty, key assumptions, and the decisions the model is intended to inform. Not all models include formal uncertainties; when they exist, report them as ranges rather than precise values to avoid implying false precision.	6 (1.25)	6 (1)	47	Consensus
C02	Uncertainty arising from model structure, assumptions, and data should be clearly specified and, where possible, quantified. It is important to specify the types of uncertainty that matter, such as parameter uncertainty, data uncertainty, and uncertainty from model assumptions. Uncertainty should be communicated alongside model results and linked to the model's intended use, clarifying which decisions it supports and how robust those conclusions are.	6 (0)	6 (0)	47	Consensus
C03	Forecasts and scenarios carry different forms of uncertainty and should be communicated differently. Forecasts involve predictive uncertainty about what will happen, while scenarios explore what could happen under specified assumptions. These represent points on a continuum rather than a strict dichotomy.	6 (2)	6 (0.75)	46	Consensus

continued on next page

Supplementary Table 4: *(continued)*

ID	Statement	R3	R4	n (R4)	Status
C04	Model validation should be clearly distinguished from calibration and verification, with validation defined as testing model performance on data not used for fitting, where feasible. Projections should not be assumed to be stable over time; validation processes should be continually updated with emerging empirical data, using adaptive criteria to guide when and how to update models.	5 (2)	6 (1)	43	Consensus
C05	Sensitivity analyses regarding parameter choices and the omission of potential other factors should be conducted with transparent methods and clearly justified parameter ranges. Initial criteria should be specified where possible but should be revisable as understanding evolves. Results should accompany all policy-informing model outputs and be communicated in accessible, decision-relevant formats.	5 (2)	6 (1)	48	Consensus
C06	Interactive visualization tools and dynamic dashboards can improve comprehension of model scenarios and limitations, but their design should be tailored to the intended audience, with attention to the risks of oversimplification and misinterpretation. They should be co-developed with users to ensure relevance and usability, regularly updated to reflect new evidence, and designed to make assumptions and uncertainties transparent. Adequate resources must be allocated for development and maintenance.	6 (2)	6 (1)	48	Consensus

continued on next page

Supplementary Table 4: *(continued)*

ID	Statement	R3	R4	n (R4)	Status
C07	Data formats and APIs should change only when necessary to capture new epidemiological variables or phenomena. When changes occur, they should be managed responsibly with clear documentation, versioning, and backward compatibility where feasible.	6 (1)	6 (0)	46	Consensus
C08	Public data websites should offer free API access, and the underlying raw or processed data should be published alongside visualizations in a machine-readable format such as CSV or JSON. When data cannot be publicly released for privacy or legal reasons, de-identified or aggregated data should be provided in a machine-readable format, accompanied by metadata and documentation to enable reuse and reproducibility.	6 (1)	6 (0)	48	Consensus
C09	Policymakers do not need technical modeling expertise, but should have a foundational understanding of model uncertainty and its implications to ensure policies align with scientific evidence. The primary responsibility for making uncertainty accessible lies with modelers and scientists, who should communicate it in accessible terms to non-expert audiences. Engaging policymakers early helps tailor how uncertainty is presented and what information is most useful for decision-making.	5 (2)	6 (1)	48	Consensus

continued on next page

Supplementary Table 4: *(continued)*

ID	Statement	R3	R4	n (R4)	Status
C10	Communication strategies for model results should address not only policymakers but also the general public and media, as public trust in modeling was a critical challenge during COVID-19.	6 (1)	6 (1)	48	Consensus
Section D: Collaboration, interdisciplinarity, and ethics					
D01	Effective pandemic preparedness requires long-term, interdisciplinary collaboration among disciplines, including epidemiologists, clinicians, virologists, microbiologists, systems engineering and operational research experts, mathematicians, statisticians, computer and data scientists, physicists, social scientists, economists, behavioral scientists, public health practitioners, and others.	6 (1)	6 (0)	47	Consensus
D02	Global modeling cooperation should be strengthened through joint platforms (modeling hubs, joint forecasting and scenario efforts) and capacity-building in LMICs and under-resourced settings. It should be supported by baseline funding for collaboration between public health institutes and academia, and by project funding to drive innovation in modeling.	6 (0)	6 (0)	47	Consensus
D03	Cross-border data-sharing agreements should be developed and strengthened to address feasibility constraints such as data protection policies, geopolitical issues, and other barriers.	6 (2)	6 (0)	47	Consensus

continued on next page

Supplementary Table 4: *(continued)*

ID	Statement	R3	R4	n (R4)	Status
D04	Dedicated communication roles, with appropriate training in bridging scientific and policy communication, can serve as mediators between scientific modeling teams, policymakers, the media, and the general public to reduce misinterpretation and increase policy uptake.	5 (1.5)	6 (1)	47	Consensus
D05	Ethical considerations should be embedded upstream into model development, validation, and scenario communication. They should explicitly address how modeling choices influence resource allocation decisions and affect vulnerable groups. The depth of ethical review should be proportionate to the decision context and urgency.	5 (2.25)	6 (1)	44	Consensus
D06	Biases related to socio-economic status, structural disadvantage, and marginalized communities should be identified and mitigated. A flexible ethical framework for guiding pandemic modeling should be developed during preparedness phases, with the understanding that it will need to be adapted as pandemic-specific challenges emerge.	5 (1)	6 (0)	45	Consensus
D07	The well-being of children and other vulnerable groups matters. Children often lack a strong advocacy voice, and their needs and interests are at times overlooked. Models informing pandemic policy should explicitly consider age-stratified impacts across vulnerable populations, recognizing that many interventions can support both infection control and continuity of education simultaneously.	5 (2)	6 (1)	47	Consensus

continued on next page

Supplementary Table 4: *(continued)*

ID	Statement	R3	R4	n (R4)	Status
D08	There is value in incorporating the lived experience of affected communities into policy making and pandemic response decisions. This should be achieved through concrete mechanisms such as community advisory boards or structured consultation, established during preparedness phases.	6 (2)	6 (1)	45	Consensus
D09	Good science communication should be supported by institutions and public funding, rather than resting solely on individual scientists. Structured exchange between science journalists and modelers can improve accuracy and policy relevance and should be encouraged, with science journalism striving to avoid sensationalism and misinformation.	6 (1)	6 (0.5)	47	Consensus

Supplementary Note 5: Top-two share and a completers-only sensitivity analysis

We defined consensus a priori as IQR ≤ 1, the most widely used dispersion criterion in Delphi practice and the one the methodological literature recommends over percentage-agreement thresholds, which are more sensitive to where the cut is drawn.[1–3] That criterion measures agreement among panellists rather than the strength of their endorsement, and two features of the Round 4 design invite a check on it. First, 38 of the 40 medians sit at the maximum of the six-point scale, so the median and the interquartile range stop separating the statements from one another, and IQR ≤ 1 becomes easy to meet. The scale carried no neutral midpoint, so a statement could in principle satisfy the criterion while sitting on the disagreeing side of it; R3-B07 at Round 3 (median 3.5) shows that this is not merely hypothetical, although no Round 4 statement comes close. Second, we retained every response that carried at least one rating, whether or not the panellist submitted it, so be-

tween 43 and 51 raters stand behind each statement while the reported panel size is the 41 completers.

We address the first by reporting the top-two share, the percentage of raters choosing 5 or 6. We report it as a post hoc robustness check on the pre-specified criterion, not as a second criterion: no statement is classified by it. It still discriminates where the median and the IQR do not: it runs from 77.1% for C09, which asks policy advisors to understand model uncertainty, to 100% for D02 on global modelling cooperation, with a mean of 92.1%. Thirteen statements fall strictly below 90%, and A06 sits at exactly 90%. Five of the thirteen (B02 and B03 at 89.8%, B04, B05, and C10 at 89.6%) round to 90%, so Supplementary Table 5 reports the share to one decimal place and the count can be checked there. The ceiling is therefore real but not total, and the statements it separates are substantively the ones the Discussion treats as contested.

We address the second by recomputing all statistics using only the 41 completers, which drops every unsubmitted response, including the six that answered all 40 items. No statement changes consensus class; all 40 remain within IQR ≤ 1. One median moves: A10 from 5.5 to 5, leaving it level with A11 as the joint-lowest of the set and not altering its classification. The count of statements at a median of 6 is unchanged at 38, and the number with an IQR of 0 rises from 13 to 16, so dispersion among completers is marginally lower rather than higher. The three sparsest statements behave the same way: A11 ($n = 45$ all raters, 37 completers), C04 (43 and 37), and D05 (44 and 39) each keep their median and their consensus class. Sparse ratings and weak endorsement coincide only for A11, whose top-two share of 80.0% is the second-lowest in the set; C04 and D05 sit at 97.7% and 90.9%.

Supplementary Table 5 reports both analyses per statement. Quartiles use linear interpolation throughout, the default in NumPy and equivalent to type 7 in R, which is why medians and interquartile ranges can take non-integer values. The analysis is reproduced by `scripts/r4_sensitivity.py` in the deposited code.

Supplementary Table 5: **Completers-only sensitivity analysis and the top-two share.** Round 4 statistics on all raters, as reported in the main text, beside the same statistics restricted to the 41 panellists who completed the round. Quartiles use linear interpolation; the top-two share is the percentage of raters choosing 5 or 6, which ranges from 77% to 100% and still separates statements where median and IQR no longer do. No statement changes consensus class, and one median moves (* A10, 5.5 to 5).

	All raters				Completers only			
ID	Median	IQR	Top-two share (%)	n	Median	IQR	Top-two share (%)	n
Section A: Model design and complexity								
A01	6	0	98.0	51	6	0	97.6	41
A02	6	1	98.0	51	6	1	97.6	41
A03	6	0	98.0	50	6	0	100.0	41
A04	6	1	92.0	50	6	1	95.1	41
A05	6	0	96.0	50	6	0	97.6	41
A06	6	1	90.0	50	6	1	92.7	41
A07	6	1	95.9	49	6	1	97.5	40
A08	6	1	93.9	49	6	1	95.0	40
A09	6	1	93.9	49	6	1	92.5	40
A10*	5.5	1	92.0	50	5	1	92.7	41
A11	5	1	80.0	45	5	1	86.5	37
A12	6	1	85.7	49	6	1	87.8	41
A13	6	0	91.8	49	6	0	95.1	41
Section B: Data availability and quality								
B01	6	0	95.9	49	6	0	97.6	41
B02	6	1	89.8	49	6	1	95.1	41
B03	6	1	89.8	49	6	0	92.7	41
B04	6	1	89.6	48	6	1	95.1	41
B05	6	1	89.6	48	6	1	90.2	41
B06	6	1	87.2	47	6	1	90.2	41
B07	6	1	91.7	48	6	0	92.7	41
B08	6	0	93.5	46	6	0	95.0	40
Section C: Uncertainty, validation, and communication								
C01	6	1	93.6	47	6	1	95.1	41
C02	6	0	95.7	47	6	0	97.6	41
C03	6	0.75	97.8	46	6	0.25	100.0	40
C04	6	1	97.7	43	6	1	97.3	37
C05	6	1	93.8	48	6	1	92.7	41
C06	6	1	85.4	48	6	1	90.2	41
C07	6	0	97.8	46	6	0	100.0	39
C08	6	0	91.7	48	6	0	95.1	41

continued on next page

Supplementary Table 5: *(continued)*

	All raters				Completers only			
ID	Median	IQR	Top-two share (%)	n	Median	IQR	Top-two share (%)	n
C09	6	1	77.1	48	6	1	80.5	41
C10	6	1	89.6	48	6	0	92.7	41
Section D: Collaboration, interdisciplinarity, and ethics								
D01	6	0	95.7	47	6	0	95.1	41
D02	6	0	100.0	47	6	0	100.0	41
D03	6	0	97.9	47	6	0	100.0	41
D04	6	1	87.2	47	6	1	85.4	41
D05	6	1	90.9	44	6	1	92.3	39
D06	6	0	95.6	45	6	0.25	95.0	40
D07	6	1	85.1	47	6	1	85.4	41
D08	6	1	84.4	45	6	1	87.2	39
D09	6	0.5	95.7	47	6	0	97.6	41

Supplementary Note 6: Round 2 to Round 3 statement revision mapping

Supplementary Table 6 summarises the changes made to Delphi statements between Round 2 and Round 3, based on the consolidated qualitative feedback from 11 contributors, namely the five Round 2 respondents and the six members of the infoXpand consortium group who conducted Round 1.

Round 3 to Round 4: the two merges Two pairs merged after Round 3, reducing the 42 statements rated in Round 3 to the 40 rated in Round 4. Round 3 statements A-2 and A-3 became A02, and C-6 and C-11 became C06. Where Supplementary Table 4 and Extended Data Fig. 1 give a Round 3 figure for A02 or C06, that figure pools the distributions of the two source statements. That is a Round 3 reading for a statement no one rated in Round 3, not a measurement of the merged item. The other 38 statements map one-to-one: Section A identifiers shift by one from A03 onwards (A03 was A-4, A11 was A-12, A13 was A-14), Section C does not shift, because the merge absorbed C-11, the last Section C statement, so C-7 to C-10 map straight onto C07 to C10, and Sections B and D are unchanged.

The two Round 3 numbering systems Round 3 carries two identifier schemes, and they are not interchangeable. Supplementary Table 6 uses sequential positions running across all sections from 1 to 42, whereas Supplementary Note 7 uses section-letter identifiers such as R3-B07. The two convert directly: positions 1 to 14 are A-1 to A-14; positions 15 to 22 are B-1 to B-8; positions 23 to 33 are C-1 to C-11; and positions 34 to 42 are D-1 to D-9. Position 27 in the table, the merged sensitivity-analysis statement, is therefore R3-C05, and position 39 is R3-D06.

Supplementary Table 6: Summary of statement revisions from Round 2 (26 statements) to Round 3 (42 statements), by section and type of change. R2 and R3 numbers are sequential positions within each round, running across all sections, so they differ from the section-letter identifiers (such as R3-B07) used in Supplementary Note 7.

R2	R3	Sec.	Change	Summary
Section A: Model design and complexity				
1	1, 2, 3	A	Split	Compound statement separated: (1) quality is question-dependent; (2) no single model fits all; (3) a limited set can address multiple questions via adaptation.
2	4	A	Reworded	Convergence now requires independent construction; divergences illuminate uncertainty.
3	5	A	Reworded	"In the best case" replaced with "In favorable conditions"; assumptions must be communicated.
4	6, 7, 8	A	Split	Split into: speed vs. uncertainty (6), complexity trade-off (7), gradual development with evaluation moments (8).
5	9, 10	A	Split	Split: parameter identifiability (9); simplified models with explicit scope and validation (10).
6+17	27	C	Merged	Overlapping sensitivity analysis statements merged; "carefully" replaced with predefined criteria.
7	11	A	Reworded	Contested claim of "less accurate predictions" removed; reframed as increased data demands requiring validation.

continued on next page

Supplementary Table 6: *(continued)*

R2	R3	Sec.	Change	Summary
8	12	A	Reworded	ABM strengths highlighted; vague cross-reference removed.
9	13	A	Reworded	Confusing negation removed; outcomes broadened to hospitalisations, deaths, long-term consequences.
—	14	A	New	Calibration reporting: method, parameters, data sources, uncertainties.
Section B: Data availability and quality				
10	23	C	Moved	Expanded to include uncertainty context, assumptions, intended decisions; moved to Section C.
11	15	B	Reworded	Quality trade-offs must document implications; preliminary data permitted with explicit caveats.
12	16, 17	B	Split	FAIR sharing (16) separated from open sharing limitations with mitigations (17).
13	18	B	Reworded	Ethical safeguards clause moved to new Statement 22; streamlined to data collection focus.
—	19	B	New	Non-traditional data sources (digital platforms, mobility, search queries, wastewater).
14	20, 21	B	Split	Hospital-entry PCR bias addressed; sentinel/population surveys reframed as complementary (20); community settings and wastewater added (21).
—	22	B	New	Ethical frameworks for data collection; inequity, privacy, predefined safeguards.
Section C: Uncertainty, validation, and communication				
15	24	C	Reworded	Specifies types of uncertainty; links to model's intended use.
—	25	C	New	Distinguishes forecast from scenario uncertainty; different communication strategies required.

continued on next page

Supplementary Table 6: *(continued)*

R2	R3	Sec.	Change	Summary
16	26	C	Reworded	Distinguishes calibration, validation, verification; adds predefined update criteria.
18	28	C	Minor edit	Core message retained; design concerns addressed in new Statement 33.
19	29	C	Reworded	"As rarely as possible" replaced with conditional language; documentation and versioning required.
20	30	C	Reworded	Blanket ban on pixel formats replaced; data published alongside visualisations.
21	31	C	Reworded	Reversed: foundational understanding now deemed important; modellers bear primary communication responsibility.
—	32	C	New	Communication strategies for the general public and media; public trust during COVID-19.
—	33	C	New	Interactive tools co-developed with users; avoid oversimplification.
Section D: Collaboration, interdisciplinarity, and ethics				
22	34	D	Reworded	Discipline list expanded (systems engineering, OR, mathematicians, statisticians, computer/data scientists, physicists).
23	35, 36	D	Split	Capacity building in LMICs with funding distinction (35); cross-border data sharing with feasibility caveats (36).
24	37, 42	D	Split	Mediator role (37); institutional responsibility and journalist-modeller collaboration (42).
25	38, 39	D	Split	Ethics "embedded upstream" (38); bias mitigation with pre-pandemic agreement on values (39).
26	40	D	Reworded	Tone aligned; child well-being framed with explicit infection-control trade-offs.

continued on next page

Supplementary Table 6: *(continued)*

R2	R3	Sec.	Change	Summary
—	41	D	New	Incorporating lived experience of affected communities.

Supplementary Note 7: Round-by-round qualitative feedback

This section provides the detailed qualitative feedback themes summarised in the results.

Round 1 (internal review)

The internal review ($N = 6$) focused on statement clarity and completeness. Panellists emphasised the question-dependent quality of models, the tension between complexity and calibration (particularly for agent-based models), and the practicality of sentinel sampling over complete population surveys. One pathogen-specific statement was deleted as overly context-dependent. Communication framing was revised: rather than requiring policymaker training, statements shifted the burden to modellers. Panellists identified several missing topics, including multi-model integration, the role of simple mechanistic models, the curse of dimensionality, sensitivity analysis as standard practice, and the stability and machine readability of published data, which together generated ten new statements. One of the ten came from a single panellist who raised children's well-being as a neglected concern. Revision against this feedback took the 17 Round 1 statements to the 26 circulated in Round 2: one deletion, one statement split in two, one pair merged into one, and those ten additions (Supplementary Note 16). Adding items in response to expert input is an accepted Delphi feature, which treats the reviewed literature as a structured starting point paired with expert judgement.[4]

Round 2 (external panel feedback)

External panellists ($N = 10$ invited) identified several compound statements that conflated distinct claims and requested that they be split. They contested the causal assertion that integrating socio-economic factors reduces

accuracy, leading to rewording reflecting conditional impacts. Vague prescriptions were made operationally specific with predefined criteria and documentation standards. Communication statements were expanded to distinguish policymakers, the public, and journalists as audiences. New statements addressed ethical frameworks, capacity building in LMICs, and the lived experience of affected communities.

Round 3 (rating and feedback)

Round 3 rated 42 statements, which the two later merges reduced to the 40 reported elsewhere, so the identifiers here carry an `R3-` prefix and do not generally match the final ones: R3-A12, for example, is the agent-based model statement that became A11. Supplementary Note 6 gives the Round 2 to Round 3 numbering.

Nine cross-cutting themes emerged from the free-text comments. The four most contested are described first, followed by the remaining five.

Agent-based model controversy The statement on agent-based models (ABMs, R3-A12) generated the most contested debate within Section A (mean 4.62, IQR 1.25). Multiple panellists challenged the claim that ABMs are "generally less suitable for short-term prediction," citing successful real-time ABM deployments during COVID-19. Others questioned the sharpness of the distinction between ABMs and compartmental models, noting that extended compartmental models can also capture substantial heterogeneity.

Sampling strategy disagreement Statement R3-B07, on community-setting sampling versus hospital-entry sampling, was the lowest-rated statement, with ratings centred on the scale midpoint (median 3.5, mean 3.75) rather than in agreement. Panellists challenged the framing that community sampling is inherently less biased than hospital sampling, noting that airports and train stations are themselves convenience samples. Multiple panellists referenced large-scale community prevalence surveys that sample the general population independently of symptoms (e.g., the Office for National Statistics COVID-19 Infection Survey and the Imperial College REal-time Assessment of Community Transmission (REACT) study) as the gold standard, rather than hospital- or convenience-community-based sampling.

Children and vulnerable populations Statement R3-D07 on children's well-being did not reach consensus at Round 3 (median 5). Panellists who rated it lower objected to singling out children rather than addressing all vulnerable groups within a unified framework. Several suggested broadening the statement to cover children and other vulnerable populations, reframing it in terms of what models should incorporate (age-stratified impacts) rather than as an advocacy statement. We revised it accordingly for Round 4, covering children and vulnerable groups together (D07), and reached consensus.

Lived experience of affected communities Statement R3-D08 showed the most polarised response distribution in Section D (mean 4.88, median 6, IQR 2), with 10 panellists rating it 6 but 4 rating it 2 or 3. The objection was not to the principle but to its wording. Several panellists called "there is value" too vague to act on, asking which value and at what trade-off, and one argued the statement was about engagement rather than about modelling and would be more pertinent if it specified how modellers should involve these communities during model development. Suggested mechanisms included community advisory boards and participatory modelling workshops. Others doubted that such engagement could be organised well and quickly enough during a pandemic.

Precision of terminology Panellists consistently requested clearer definitions of key terms. The word "quality" in Statement R3-A01 was flagged as ambiguous, potentially referring to fitness for purpose, technical correctness, or predictive accuracy. The term "curse of dimensionality" (R3-A09) was considered jargon by some panellists from non-mathematical backgrounds. "Favorable conditions" (R3-A05) and "information completeness" (R3-A07) were similarly flagged as vague.

Tensions between aspiration and feasibility A recurring theme across sections was the tension between ideal practices and what is achievable during a crisis. Panellists noted that formal uncertainty quantification (R3-A06) can be computationally prohibitive under time pressure and that qualitative descriptions of uncertainty may need to suffice initially. Similarly, predefined evaluation moments for model development (R3-A08) and predefined criteria for sensitivity analyses (R3-C05) were criticised as too rigid for rapidly evolving pandemic contexts. Several panellists preferred "iterative" or "adaptive"

language over "formal" or "predefined."

Data sharing: principle versus practice While the principle of open data sharing received strong endorsement (R3-B02: mean 5.41, IQR 1), the companion statement on limitations of open sharing (R3-B03: mean 4.47, IQR 1) revealed tension. Panellists noted that data access agreements and synthetic data generation are too slow for emergency response and that privacy concerns are sometimes overweighed relative to public health needs. Several panellists argued that aggregate-level data (e.g., hospital admissions) poses minimal privacy risk and should be shared openly by default.

Communication responsibility Across Sections C and D, panellists debated the division of communication responsibility between modellers and policymakers. While there was broad agreement that the primary burden falls on modellers to communicate accessibly (R3-C09: mean 5.06, median 5), panellists noted that some policymakers lack the willingness or capacity to engage with model-based evidence regardless of how it is presented. The distinction between communicating to policymakers, the general public, journalists, and interdisciplinary teams was repeatedly raised, with panellists calling for differentiated strategies for each audience.

Ethics: upstream versus downstream The ethics statements (R3-D05, R3-D06) generated substantial discussion about whether ethical considerations belong in model development itself or only in how model outputs are applied to decisions. Some panellists argued that mathematical models are inherently value-neutral tools and that ethics enters only at the application stage. Others countered that choices about which populations to model, which outcomes to prioritise, and which scenarios to explore embed values that warrant ethical scrutiny. The requirement to articulate ethical values before a pandemic (R3-D06: mean 4.71, IQR 1) was widely considered impractical, since each pandemic presents unforeseen challenges.

Round 4 (final rating): section comments and open-ended responses

The section-level comments mostly endorsed the statements while adding nuance, and several themes recurred across sections. In Section A (29 com-

ments), panellists reaffirmed the parsimony principle and stressed that simplicity and transparency must be balanced against the need to capture key dynamics, especially under crisis-time resource constraints. In Section B (21 comments), they argued that data governance frameworks, metadata standards, and pre-negotiated sharing agreements should be in place before a pandemic. They cautioned that ambitious data collection must respect privacy and address unequal national capacities. In Section C (20 comments), they called for two-way communication in which modellers understand policy needs and policymakers build foundational scientific literacy during inter-pandemic periods, and they flagged disinformation and the need for audience-tailored messaging. In Section D (18 comments), they highlighted a fragmented data-sharing infrastructure and international inequity and urged continuous funding for interdisciplinary collaboration and science communication rather than improvisation during crises.

The seven open-ended questions each drew between 35 and 38 responses. On the most important lesson from past pandemic modelling (F-1), panellists' answers centred on building the ecosystem around models, namely, clear uncertainty communication, pre-established data pipelines, trusted modeller-policymaker relationships, and institutional preparedness, and on framing models as decision-support tools rather than precise forecasts. Asked what most models failed to address during COVID-19 (F-2), they pointed above all to dynamic human behaviour, including risk perception, compliance, policy fatigue, and socio-economic heterogeneity. On public trust (F-3), they attributed distrust to communication failures, the prevention paradox, political polarisation, and disinformation rather than to technical inaccuracy alone. On the biggest current gap (F-4), they emphasised the availability, quality, timeliness, and interoperability of data, followed by the integration of behavioural dynamics and sustained funding. In urgent collaborations (F-5), they prioritised the behavioural and social sciences and a stronger modeller-policymaker interface. In integrating models into policy (F-6), they called for standing advisory structures and bidirectional capacity building established during inter-pandemic periods. On ethics (F-7), they identified equity and transparency as paramount, urging models to account for differential effects on vulnerable populations and to state assumptions and limitations explicitly.

Supplementary Note 8: Ranking of preparedness measures

Panellists ranked seven preparedness measures across three phases: before a pandemic emerges, during the early stage of emergence, and after the event (N = 45–46 per phase). Supplementary Table 7 reports every measure in every phase, with the first-choice count, the mean rank, and the Borda total. This task asked the panel to state importance directly, which the statement ratings do not measure: those record agreement, and a statement can command consensus yet matter little for preparedness. The seven measures do not correspond one-to-one to the ten priorities in Table 3 of the main article.

Real-time data sharing and *rapid-response modelling teams* were the top two measures for both the pre-pandemic and early-emergence phases: *real-time data sharing* ranked first before a pandemic (mean rank 2.06, first choice for 22 of 46 panellists), and the two were effectively tied at the top during early emergence (mean ranks 2.07 and 2.04). The ordering shifted in the post-event phase, with *empirical model validation* moving to first place (mean rank 3.09), followed closely by *socio-behavioural science inclusion* and *real-time data sharing*, while *rapid-response modelling teams* fell to last (mean rank 5.52). That post-event ordering is the least settled of the three. The leading measures' ranks span an IQR of 4, and those of *real-time data sharing* span 5, against 1.75 for the pre-pandemic leader, so the means there separate measures the panel itself did not consistently separate. *Science communication tools* and *ethical frameworks for decision-making* never reached the top three in any phase, though neither sat consistently low: ethical frameworks placed fourth of seven both before a pandemic (mean rank 4.24) and after the event (3.75), and science communication tools placed fourth during early emergence (4.53). The panel therefore prioritised data infrastructure and rapid-response capacity for the pre-pandemic and early-emergence phases and methodological consolidation through validation and behavioural integration for recovery and preparedness.

Supplementary Table 7: **Ranking of preparedness measures by pandemic phase.** The Round 4 ranking task, in which panellists ordered the same seven measures by importance for each of three phases, rank 1 being the most important. Measures are listed by mean rank within each phase. "First" counts the panellists who ranked a measure first. Borda awards $8 - r$ points for rank r, so a higher total means a measure was preferred. Because it is a sum rather than an average, it can order two measures differently from the mean rank when their n differ. The IQR is the interquartile range of the ranks a measure received, so a small value means panellists placed it consistently and a large one that they did not. N is the number of panellists who ranked a phase at all, and varies across phases because a panellist could rank one phase and leave another blank; n is the number who ranked that particular measure. This task measured stated importance, unlike the statement ratings in Supplementary Table 4, which measure agreement.

Preparedness measure	n	First	Mean rank	IQR	Borda
Before an epidemic or pandemic emerges ($N = 46$ panellists)					
Real-time data-sharing systems	46	22	2.06	1.75	273
Development of rapid-response modeling teams	45	10	3.33	2.00	210
Definition and usage of standardized modeling protocols	45	6	4.11	4.00	175
Ethical frameworks for decision-making based on model outcomes	46	4	4.24	3.00	173
Enhanced validation of models with empirical data	45	2	4.36	3.00	164
Greater inclusion of socio-behavioral sciences in models	45	0	4.76	2.00	146
Improved science communication tools (e.g., interactive dashboards, uncertainty visualizations) for policymakers	46	2	5.02	2.75	137

continued on next page

Supplementary Table 7: *(continued)*

Preparedness measure	n	First	Mean rank	IQR	Borda
During the early stage of emergence ($N = 45$ panellists)					
Development of rapid-response modeling teams	45	19	2.04	1.00	268
Real-time data-sharing systems	45	20	2.07	1.00	267
Enhanced validation of models with empirical data	45	2	4.20	3.00	171
Improved science communication tools (e.g., interactive dashboards, uncertainty visualizations) for policymakers	45	1	4.53	2.00	156
Definition and usage of standardized modeling protocols	45	0	4.91	2.00	139
Ethical frameworks for decision-making based on model outcomes	45	3	5.02	3.00	134
Greater inclusion of socio-behavioral sciences in models	45	0	5.22	2.00	125
After the event, in preparation for potential future pandemics ($N = 45$ panellists)					
Enhanced validation of models with empirical data	44	15	3.09	4.00	216
Greater inclusion of socio-behavioral sciences in models	44	3	3.50	2.25	198
Real-time data-sharing systems	45	14	3.58	5.00	199
Ethical frameworks for decision-making based on model outcomes	44	3	3.75	3.00	187
Definition and usage of standardized modeling protocols	45	6	4.07	4.00	177
Improved science communication tools (e.g., interactive dashboards, uncertainty visualizations) for policymakers	44	4	4.39	3.00	159
Development of rapid-response modeling teams	44	0	5.52	2.00	109

Supplementary Note 9: Frontline services as a data and implementation interface

Frontline clinical and public health services deserve explicit attention as both data sources and delivery points for model-informed policy. Which service sits on the front line depends on the setting. In England, family physicians generated frontline COVID-19 data and implemented model-informed guid-

ance through the National Health Service,[5] which is one expression of the role primary care already holds in a health system.[6,7] Panellists described a different pattern where incidence stayed low for long periods, as in New Zealand: primary care was not the frontline in the same way, and local public health units were. Those teams held detailed working knowledge of contact patterns, behaviour, and how interventions played out in practice, and little of it ever reached a dataset or a model. Either way, the pattern is the same: the services closest to the outbreak both feed models and act on them, and the modelling literature rarely makes that loop explicit.

Three implications follow, and they apply to whichever service holds the frontline role in a given system. Its information systems should interoperate with real-time surveillance so that routine operational data reaches models with minimal delay (B01, C07, C08). Modelling groups should include practitioners from that service (D01). And policy-facing models should be validated against the incidence signal generated by frontline services,[8] which is independent of the reporting streams that models typically rely on (C04).

More established guidance in related fields offers templates this field could adapt, such as the Overview, Design concepts, and Details (ODD) protocol for agent-based models[9,10] and the STRengthening Analytical Thinking for Observational Studies (STRATOS) initiative for statistical reporting.[11]

Supplementary Note 10: Recommendations by audience

We assigned a party when the statement text places a duty on it, or when the modelling group cannot enact the statement without it, and not when a party appears only as the recipient of communication. Several statements ask modellers to explain their work to policymakers (A04, C01, C02, C03), and those carry M alone, since the modellers must act. A03 likewise carries M alone: a modelling group can compare its own model against structurally different ones already published, so the insight it asks for needs no agreement with the groups that built them. A statement whose initiator is an agency carries A even where modellers are among the participants, as in the standing collaborations of D01 and the advisory mechanisms of D08, and a statement addressed to modelling practice carries M even where the practice it asks for is field-wide, as in D06. The parties differ in what they must do. The identifiers in the list below illustrate each kind of duty rather than enumerating every statement that carries it; Table 1 of the main article holds the complete labelling.

- **Modelling groups** bear the burden of reporting and communicating uncertainty (A13, B02, C01, C02, C03, C05, C09, C10).
- **Public health agencies** build the relationships, pipelines, and advisory structures that have to exist before a crisis (B01, D01, D02, D04).[12]
- **Policy advisors** need no technical expertise but do need a working sense of what models can deliver (C09), and they share responsibility for the ethical safeguards agreed before a pandemic (B08) and for cross-border data-sharing agreements (D03).
- **Infrastructure providers** supply the interoperable, privacy-preserving platforms that the rest depends on (B01, B02, B03, C07, C08, D03).

Recommendations written largely by modellers, 22 of which need someone beyond the modelling group to act, can read as a wish list addressed to someone else. In one sense, they are: they describe an arrangement we consider ideal, not one negotiated with the parties who would have to build it, and this panel cannot speak for those parties. These recommendations say what modelling groups need from public health agencies. The counterpart we cannot write alone says the reverse: what agencies need from modelling groups, on what timescale, and in what form. Side by side, the two would expose where expectations conflict, which neither can see alone.

Two of the ten priorities in Table 3 of the main article are also constrained by when they can be delivered. *External model validation* relies on data from after the outbreak, which may be why the panel ranked it a top-priority measure for that phase. *Behavioural and social science integration* imposes additional data and calibration demands, and collecting reliable behavioural data requires resources and time, so such data remains scarce. Both are constrained by when they can be delivered, not by whether they are worth doing.

Supplementary Note 11: Expert panel recruitment and composition

We recruited the expert panel through the infoXpand consortium, a network of researchers with expertise in infectious disease modelling, transmission

dynamics, public health, science-policy advisory, and the social sciences. We selected panellists based on demonstrated expertise in at least one of these domains, evidenced by peer-reviewed publications, participation in pandemic advisory bodies, or active involvement in modelling for public health decision-making.

We recruited through personalised email invitations, first to consortium members and then, in Rounds 2 to 4, to external experts identified through professional networks. Continued recruitment through those networks took the invitation list from 38 experts in Round 3 to 62 in Round 4; the additional invitees were identified the same way as their predecessors, by publication record, advisory role, or direct involvement in modelling for policy.

Round 4 differed from the earlier rounds in one further respect. We distributed it as a shareable link rather than as individually keyed invitations, and we encouraged panellists to forward it to colleagues they judged qualified. The link carried no per-recipient key, and the round collected no identifying data, so we cannot say how many respondents arrived that way or whether any did. Several panellists told us they had forwarded it, and that is all we know. Two consequences follow. First, the denominator is not fixed: more people saw the survey than the 62 we invited directly, so the 41 complete responses cannot be converted into a response rate. Fig. 3 of the main article reports the two counts separately rather than as a ratio. Second, we did not screen the respondents who arrived by onward sharing. Consent was unaffected, since the survey itself opened with the consent statement every respondent affirmed before rating, and the round was anonymous and collected no personal data. Eligibility, however, rested on panellists applying the criteria we gave them, and the self-reported characteristics of the final panel are consistent with that having worked. Of the 41 completers, 40 are academic researchers and 39 of those work in universities or research institutes, 40 published on the topic within the past five years, and 36 report five or more years of relevant experience, 27 of them more than a decade. The same table also includes figures that run counter to this: six panellists rate their own expertise as of some or limited depth, and five report fewer than five years of relevant experience. None of this verifies eligibility on a case-by-case basis, and we treat it in the main article as a limitation rather than a control.

One consequence deserves stating plainly. As Supplementary Note 10 sets out, 22 of the 40 statements need a public health agency, a funder, a policy advisor, or an infrastructure provider to act, yet nobody in any of those

roles rated them: 40 of the 41 completers are academic researchers. The parties on whom the set places duties did not vote on it. The gap applies with particular force to D02, on capacity building in low- and middle-income countries, which a panel drawn overwhelmingly from Europe, with only 2 of the 41 in North America, endorsed unanimously (100% top-two share).

We explicitly considered geographic and disciplinary diversity during recruitment, although the resulting panel is concentrated in Europe, reflecting the consortium's institutional links. We also included experts from Australia and New Zealand, who were central to their countries' COVID-19 response efforts. This limits the scope to high-income settings with comparable institutional and technical capacity.

Two groups need to be distinguished. The byline note for the main article lists everyone who took part in at least one Delphi round, which is the broader contributor group. Throughout the article, "the panel" means the 41 respondents who completed the definitive Round 4 rating. Because that round was anonymous, the two cannot be matched to one another: we can say how many completed it, not which of the named contributors did. Extended Data Table 1 summarises the 41 panellists who completed the final Round 4 rating. The panel was predominantly based in Europe and almost entirely composed of academic researchers in universities or research institutes. It skewed towards senior career stages, with most respondents holding professorial or senior research positions and reporting more than a decade of relevant experience. Nearly all had published peer-reviewed work or contributed to formal guidelines on the topic within the past five years.

The final-round panel is narrower than the study's wider contributor group, which spans 22 countries across five continents, two-thirds of whom report more than a decade in the field, and several of whom have advised on outbreak response abroad. We do not offer that breadth as a substitute for representation on the panel that produced the ratings, and the main article treats the composition of the final-round panel as a limitation of the study.

Supplementary Note 12: Relevance framework

The first version of this framework drew on AMSTAR 2,[13] guidance on matching the type of review to the question it asks,[14] and STROBE.[15] It scored publications against six criteria, each on a scale of 0 to 2, for a maximum of 12, and classified them as high (10 to 12), moderate (6 to 9), or low (below 6) in relevance. Two of those criteria, public health relevance and

modelling relevance, restated eligibility rather than measuring relevance: the first scored the maximum for every publication, and the second nearly did. The framework placed 96% of publications in the top band and rated 21 of the 24 publications that the sharpened criteria excluded as highly relevant. We therefore retired it, moved eligibility into the five criteria of Supplementary Note 13, and rebuilt the relevance framework around the four dimensions in Supplementary Table 8. Its anchors state what each score requires, so the middle option must be earned rather than defaulted to. The rebuilt framework nonetheless still places 92 of the 135 publications in its top band, 40 in the moderate band, and 3 in the low band, so it discriminates more than the 96% of its predecessor but not sharply. No result reported in the article uses these bands: nothing is weighted, stratified, or restricted by relevance, and the consensus results do not depend on them. We report them as a description of the corpus and nothing more.

Supplementary Table 8: Relevance framework criteria and scoring rubric. Each criterion is scored 0 to 2, yielding a maximum total of 8. Publications are classified as high (7-8), moderate (5-6), or low (below 5) in relevance. Eligibility is assessed separately against the five criteria in Supplementary Note 13.

Criterion	**0**	**1**	**2**
Evidence base	Substantially unreferenced assertion. Few or no citations supporting the central claims.	Partly referenced. Key claims are cited, but others rest on assertion or authority, or the citation set is thin, one-sided, or heavily self-referential.	Claims are consistently supported by citations to primary studies, reviews, or seminal work. Where the authors advance a contested position, they engage with the literature on both sides.
Breadth of discussion	Confined to a single narrow issue or a single aspect of modelling.	Covers more than one aspect but stays within one register, such as several technical issues, or touches broader themes only in passing.	Spans several distinct dimensions, such as technical, data, institutional, communication, and ethical, and relates them to one another rather than listing them.
Acknowledgement of uncertainty and bias	Little or no acknowledgement, and sustained one-sided advocacy.	Uncertainty or limitations are acknowledged but handled briefly or formulaically, and the paper leans clearly towards one position without seriously engaging alternatives.	Substantive treatment of uncertainty, limitations, or bias, with a balanced view that credits competing positions or acknowledges where the authors' own stance may be wrong.
Transparency	Opaque. Conclusions cannot be traced to a described process, and no relevant disclosure is offered.	Partially transparent, with material gaps remaining, such as a review that names databases but not inclusion criteria.	Fully transparent for its type. Reviews describe search strategy, selection, and synthesis. Research articles give methods, data, and assumptions sufficient to replicate. Opinion pieces and editorials disclose the authors' background, role, or competing interests and make the basis for their claims clear.

Supplementary Note 13: Inclusion and exclusion criteria

Beyond the five content criteria below, we restricted eligibility to publications available in English. The database queries had no language filter, so non-English records were included in the search and excluded at screening. We did not log language as an exclusion reason on its own, so those records are included in the screening counts in the PRISMA diagram, and we cannot report how many were excluded on that basis alone.

We included a study if it met all five of the following criteria:

1. **Scope.** The paper focuses on the spread of human infectious diseases in the context of pandemics, epidemics, or disease outbreaks. We excluded papers addressing only non-communicable diseases, chronic diseases, animal diseases, or non-infectious health conditions.

2. **Model-based approach.** The paper focuses on mathematical models of infectious diseases, including but not limited to ordinary differential equations, agent-based models, stochastic models, statistical models, Bayesian models, and machine learning models. We excluded papers that described only data trends without reference to mathematical models.

3. **Public health relevance.** The paper discusses how mathematical models contribute to public health decision-making, including policy-making, pandemic preparedness, interventions, and resource allocation. We excluded papers exclusively focused on theoretical model development without real-world applications.

4. **Challenges.** The paper includes a critical discussion of the role, challenges, limitations, gaps, and future research directions in infectious disease modelling. We excluded papers that presented only a model and its results, without broader reflection.

5. **Broader perspective.** The paper goes beyond presenting a single model by adopting a broader perspective on mathematical modelling, including opinion pieces, perspective articles, editorials, reviews, and general discussions. We excluded papers that presented and discussed only a single model or its technical aspects.

Criteria 4 and 5 carry most of the screening decisions, and both rest on judgements that are easy to state loosely. The notes below are the operational guidance the reassessment actually applied, condensed from the screening prompt reproduced verbatim in Supplementary Note 15.

For **model-based approach** (criterion 2), the paper must be *about* modelling, but need not contain, develop, or run a model of its own. A review, editorial, perspective, or commentary whose subject is infectious disease modelling meets this criterion fully while presenting no model at all. Criterion 5 admits exactly those formats, so requiring an original model here would contradict it. Marking a paper as not met because it does not itself develop or apply a model is the commonest error on this criterion.

For **public health relevance** (criterion 3), we score the body, not the framing. Papers routinely promise policy relevance in the abstract and deliver none. An applied forecasting study that never discusses how the forecast was or should be used does not meet this criterion, however topical its subject.

For **critical discussion** (criterion 4), a boilerplate limitations paragraph, such as a note that the model assumes homogeneous mixing, is not a critical discussion of the role and limitations of modelling. We focused on engagement with what modelling can and cannot do, rather than on a required disclosure.

For **broader perspective** (criterion 5), we applied a deletion test: if the paper's own model and results were removed, would anything remain that is a claim about modelling as a practice? Method selection rationale is the most common source of false positives. An argument that mean-field models miss clustering, so an agent-based model is used, justifies the study's own design and does not meet this criterion, however critically argued. Bibliometric and scoping reviews that map who publishes what, without examining how modelling is done, also fail here. A research article reporting original model results starts from the premise that it is not met and has to earn its way out.

Criteria 4 and 5 are related but not redundant. A paper can reflect critically on modelling in general while reporting a single model, satisfying criterion 4 but not 5, and a review can survey many models without offering any critical claim about practice, satisfying 5 but not 4. We required all five.

Supplementary Note 14: Search strings

We used the following database-specific search queries in the systematic review, all run on 7 February 2025. The generic search structure combined

four concept blocks: disease context, modelling approach, public health relevance, and critical reflection. Each query then restricted results by subject classification: MeSH terms in PubMed, index terms in Scopus, and Web of Science Categories in Web of Science, the last of these a journal-level rather than record-level filter.

The blocks themselves were adapted to each database rather than held identical, so the three queries are not translations of one another. PubMed and Scopus share the same four blocks. The Web of Science query differs in three of them: the modelling block adds `simulation*` and `data-driven model*` and narrows `stochastic model*` to `stochastic epidemic model*` and `stochastic disease model*`; the policy block is broader, adding `policymaking`, `policy support`, `policy impact`, `pandemic response`, `epidemic response`, `health policy`, and `health planning`; and the reflection block replaces the PubMed set with `improvement*`, `enhancement*`, `best practice*`, `guideline*`, `future direction*`, and `lessons learned`. We widened it because the Web of Science Categories filter is at the journal level and therefore blunter than the record-level index terms available in the other two databases.

Two limitations of the queries are worth stating. The modelling block carries the standalone term `modelling` without a wildcard and without its American variant, so a record whose only qualifying term is the bare word *modeling* is not retrieved. The wildcarded phrases (`mathematical model*`, `epidemic model*`, and the rest) match either spelling, so the gap is confined to records using no such phrase. Given how much of this literature is written in American English, we cannot rule out that it cost us relevant records. Second, all three searches ran on 7 February 2025 and we did not run an update search, so the review is current only to that date, and any record published or indexed afterwards falls outside it by construction.

Whitespace is not significant in any of these queries. Where a line runs past the margin below it is wrapped without a continuation marker, so each block can be copied as it stands.

PubMed

```
(("infectious disease*"[Title/Abstract] OR
"pandemic*"[Title/Abstract] OR "epidemic*"[Title/Abstract] OR
"disease outbreak*"[Title/Abstract]) AND
("modelling"[Title/Abstract] OR "mathematical
model*"[Title/Abstract] OR "compartmental model*"[Title/Abstract] OR
"ODE model*"[Title/Abstract] OR "ordinary differential
equation*"[Title/Abstract] OR "agent-based model*"[Title/Abstract]
OR "ABM simulation"[Title/Abstract] OR "stochastic
model*"[Title/Abstract] OR "statistical model*"[Title/Abstract] OR
"Bayesian"[Title/Abstract] OR "machine learning"[Title/Abstract] OR
"infectious disease model*"[Title/Abstract] OR "epidemic
model*"[Title/Abstract] OR "pandemic model*"[Title/Abstract]) AND
("public health"[Title/Abstract] OR
"decision-making"[Title/Abstract] OR "decision
making"[Title/Abstract] OR "policy*"[Title/Abstract] OR "pandemic
preparedness"[Title/Abstract] OR "epidemic
preparedness"[Title/Abstract] OR "resource
allocation"[Title/Abstract] OR "intervention*"[Title/Abstract]) AND
("role"[Title/Abstract] OR "challenge*"[Title/Abstract] OR
"gaps"[Title/Abstract] OR "limitation*"[Title/Abstract] OR "future
research"[Title/Abstract])) AND ("Health Policy"[MeSH] OR "Public
Health"[MeSH] OR "Health Planning"[MeSH])
```

Scopus

```
(TITLE-ABS-KEY("infectious disease*" OR "pandemic*" OR "epidemic*"
OR "disease outbreak*") AND TITLE-ABS-KEY("modelling" OR
"mathematical model*" OR "compartmental model*" OR "ODE model*" OR
"ordinary differential equation*" OR "agent-based model*" OR "ABM
simulation" OR "stochastic model*" OR "statistical model*" OR
"Bayesian" OR "machine learning" OR "infectious disease model*" OR
"epidemic model*" OR "pandemic model*") AND TITLE-ABS-KEY("public
health" OR "decision-making" OR "decision making" OR "policy*" OR
"pandemic preparedness" OR "epidemic preparedness" OR "resource
allocation" OR "intervention*") AND TITLE-ABS-KEY("role" OR
"challenge*" OR "gaps" OR "limitation*" OR "future research")) AND
INDEXTERMS("Health Policy" OR "Public Health" OR "Health Planning")
```

Web of Science

```
(TS=("infectious disease*" OR "disease outbreak*" OR "pandemic*" OR
"epidemic*") AND TS=("modelling" OR "mathematical model*" OR
"simulation*" OR "compartmental model*" OR "ODE model*" OR "ordinary
differential equation" OR "agent-based model*" OR "stochastic
epidemic model*" OR "stochastic disease model*" OR "statistical
model*" OR "data-driven model*" OR "machine learning" OR "Bayesian")
AND TS=("policymaking" OR "policy support" OR "policy impact" OR
"decision making" OR "intervention*" OR "preparedness" OR "pandemic
response" OR "epidemic response" OR "resource allocation" OR "health
policy" OR "public health" OR "health planning") AND
TS=("improvement*" OR "limitation*" OR "enhancement*" OR
"challenge*" OR "best practice*" OR "guideline*" OR "future
direction*" OR "future research*" OR "lessons learned")) AND
WC=("Health Policy & Services" OR "Public, Environmental &
Occupational Health" OR "Infectious Diseases" OR "Mathematical &
Computational Biology" OR "Epidemiology")
```

We ran no backward or forward citation searching. A title-and-abstract strategy that additionally requires a reflection term will miss publications whose reflective content is not signalled in either field, and one round of citation chasing from the included set would have partly compensated for that. We did not do it, and we note it here as a limitation of the search rather than a deliberate restriction.

Supplementary Note 15: Use of large language models

We used large language models as assistants under the authors' supervision at five points in the study, as described below. Two belong to the review: the preliminary screening, relevance scoring, and structured extraction of the retrieved articles, and the later eligibility reassessment against the sharpened criteria. Two reduce free text to candidate themes: sorting the extracted challenge and recommendation items into the eight themes of Supplementary Note 18, and synthesising the Round 3 and Round 4 panellist comments, the latter on a locally hosted model so that no panellist-authored text left institutional control. The fifth is the only generative use and the only one that touched the statements: a project-specific GPT-5 agent drafted the Round 1 set from the extracted review items and a team-authored document of candidate ideas, which the team then reviewed, consolidated, and rewrote against the reviewed literature before any panellist saw it. Every statement the panel

rated is therefore author-written, and no result was produced by a model. Separately from these five, and after the analysis was complete, the authors used Grammarly, Quillbot, and Anthropic Claude to check the language, style, and structure of the manuscript, reviewing and editing everything that those tools flagged. No passage of this manuscript or its Supplementary Information was drafted by a model. This reflects a wider move towards LLM-supported evidence synthesis, accompanied by close scrutiny of how reliably such systems perform and how their use should be reported.[16–18] We remained aware of the documented weaknesses of these models, in particular their tendency to generate fabricated or unsupported content,[19,20] their sensitivity to prompt wording, and their variability between runs. We therefore treated every LLM output as a preliminary first pass, and the authors made all substantive decisions, including inclusion, exclusion, relevance categorisation, and the final wording of the themes and statements. Systematic verification against the source articles was confined to the eligibility reassessment, during which one author (L. Stellbrink) checked 84 of the 159 reassessed records, including every proposed exclusion and a random sample of the inclusions. The outputs from the initial screen and the extraction were not systematically verified against the sources. We were unable to verify all outputs, and we did not compute a formal agreement statistic between the models and the reviewers. We note this as a limitation and a priority for future work, for example, a redesigned or independently validated scoring step.

We did not log version strings beyond the identifiers given, access dates, or sampling settings, so we cannot report them. TRIPOD-LLM asks for these,[18] and their absence is a reporting limitation rather than a judgement that they do not matter. Each record was processed once in each step.

For the initial screen of the review, we used OpenAI GPT-4o mini (model identifier `gpt-4o-mini`) to produce preliminary, structured assessments of the retrieved full-text articles. We accessed the model through the OpenAI Assistants API with the `file_search` tool, uploading each article's PDF and attaching it to a single structured query. For each paper, the model returned a JSON object in three parts: descriptive extraction fields (paper type, keywords, and the diseases, modelling approaches, social contexts, populations, geographical locations, time periods, data sources, and other factors the paper discussed); a 0–2 score with a keyword comment for each of six criteria (public health relevance, modelling relevance, evidence base, breadth of discussion, acknowledgement of uncertainty and bias, and transparency); and

bulleted lists of the challenges and recommendations it discussed. The first two criteria also served as the inclusion criteria, and all six as the relevance framework then in use. The model produced only preliminary values, and the complete screening script is available with the study materials. Eligibility itself was decided by the research team, not by those two scores. We recorded the criteria each excluded report failed, and those are the reasons Fig. 2 of the main article gives for the initial full-text exclusions.

For the reassessment, we used Anthropic Claude Sonnet 5 with the prompt reproduced below. It differs from the prompt above in two respects that matter for the results. Eligibility is assessed criterion by criterion, each requiring a quoted passage, a line of reasoning, and a verdict of "met," "not met," or "unclear," rather than folded into two relevance scores. And the four relevance dimensions are scored against stated anchors at each level, so the middle option must be earned rather than defaulted to. The prompt is exported directly from the module the screening pipeline runs, so it appears exactly as issued, with American spelling and all. The wrap markers in the left margin below are inserted by the typesetting process and are not part of the prompt.

```
You are screening papers for a systematic review on the role of
↪  mathematical
modeling of infectious diseases in public health decision-making.

For each paper you are given the full text. Read it before answering. Base
↪  every
judgement on what the attached document actually says -- not on the title,
↪  not
on the journal, and not on what you may know about the paper from
↪  elsewhere.

## 1. Inclusion criteria

The review has five criteria. Each is a requirement, so assess each one
separately as **met**, **not met**, or **unclear**. A paper is included
↪  only if
all five are met -- but do not work backwards from a decision you have
↪  already
formed: judge each criterion on its own evidence and let the decision
↪  follow.

Use **unclear** only when the document genuinely does not settle the
↪  question --
```

```
a truncated or unreadable text, or a criterion the paper never addresses either
way. It is not a hedge for a judgement you would rather not make.

For each criterion, in this order: quote the passage that most directly bears on
it, state your reasoning in one line, then give the verdict.

### Criterion 1 -- Scope

The paper must focus on the spread of infectious diseases in the context of pandemics, epidemics, or disease outbreaks.

*Excluded:* Papers that exclusively address non-communicable diseases, chronic diseases, or non-infectious health conditions are excluded.

How to apply it: Nearly every paper reaching full-text screening passes this. Mark it not met only for a genuine scope failure -- a paper about chronic disease, non-communicable conditions, or health systems with no infectious disease content.

### Criterion 2 -- Model-Based Approach

The paper must focus on mathematical models of infectious diseases, including but not limited to ordinary differential equations (ODE), agent-based models (ABM), stochastic models, statistical models, Bayesian models, and machine learning models (e.g. reinforcement learning, neural networks).

*Excluded:* Papers that exclusively describe data trends without referring to mathematical models (e.g. purely descriptive epidemiological studies, surveys, clinical studies) are excluded.

How to apply it: 'Focus on mathematical models' means the paper is *about* modeling. It does NOT mean the paper must contain, develop, or run a model of its own. A review, editorial, perspective or commentary whose subject is infectious disease modeling meets this criterion fully while presenting no model at all -- criterion 5 explicitly admits exactly those formats, so requiring an original model here would contradict it. This is the commonest error on this criterion: do not mark a paper 'not met' because it 'does not itself develop or apply a model'.
```

```
    What fails instead is a paper where modeling is not the subject: a
    ↪ descriptive epidemiological study, a survey, a clinical study, or
    ↪ a policy analysis that cites model outputs as background evidence.
    ↪ Statistical, Bayesian and machine-learning models all count, so do
    ↪ not read 'mathematical' narrowly -- but reporting fitted trends is
    ↪ not the same as modeling disease dynamics.

### Criterion 3 -- Public Health Relevance

The paper must discuss how mathematical models contribute to public health
↪ decision-making, including but not limited to policymaking,
↪ pandemic/epidemic preparedness, interventions, and resource
↪ allocation.

*Excluded:* Papers that exclusively focus on theoretical model development
↪ without discussing real-world applications or policy implications are
↪ excluded.

How to apply it: Score the body, not the framing. Papers routinely promise
↪ policy relevance in the abstract and deliver none. An applied
↪ forecasting study that never discusses how the forecast was or should
↪ be used does not meet this criterion, however topical its subject.

### Criterion 4 -- Challenges

The paper must include a critical discussion of the role, challenges,
↪ limitations, gaps, or future research directions in infectious disease
↪ modeling.

*Excluded:* Papers that exclusively present a model and its results
↪ without discussing its implications, strengths, or weaknesses are
↪ excluded.

How to apply it: A boilerplate limitations paragraph -- 'our model assumes
↪ homogeneous mixing' -- is not a critical discussion of the role and
↪ limitations of modeling. Look for engagement with what modeling can
↪ and cannot do, not a required disclosure.

### Criterion 5 -- Broader Perspective

The paper must go substantially beyond presenting a single model and its
↪ implications by taking a broader perspective on mathematical modeling,
↪ including but not limited to opinion pieces, perspective articles,
↪ editorials, reviews, and general discussions.
```

```
*Excluded:* Papers that exclusively present and discuss a single model,
↪  its results, or technical modeling aspects (e.g. parameter estimation,
↪  numerical solutions, computational efficiency, scenario-specific
↪  results) without discussing broader insights, challenges, or future
↪  directions in infectious disease modeling are excluded.

How to apply it: Apply the deletion test: if this paper's own model and
↪  results were removed, would anything remain that is a claim about
↪  modeling as a practice? Method-selection rationale is the commonest
↪  false positive -- 'mean-field models miss clustering, so we use an
↪  agent-based model' justifies the study's own design and does not meet
↪  this criterion, however critically argued. Bibliometric and scoping
↪  reviews that map who publishes what, without examining how modeling is
↪  done, also fail here. A research article reporting original model
↪  results starts from 'not met' and has to earn its way out.

## 2. Publication type

Classify by what the paper *does*, not by where it appeared or what it is
labelled. Apply in order and take the first that fits:

- Systematic review / Meta-analysis: states an explicit, reproducible
↪  search and
  selection procedure. Meta-analysis additionally pools results
  ↪  statistically.
- Review article: surveys a body of literature without a formal search
↪  protocol.
- Research article: presents original analysis, data, or model results.
- Perspective/Opinion: advances the authors' argued position, typically
↪  without
  new data. Includes "viewpoint" and "essay" formats.
- Editorial: written by journal editors or invited to frame an issue or
↪  section.
- Commentary: responds to a specific paper, event, or policy.
- Letter: short correspondence.
- Conference paper, Technical report, Policy paper, Book chapter, Book,
↪  Thesis,
  Preprint: use when the venue determines the form.
- Other: only when nothing above fits. State what it is in the evidence
↪  field.

A journal's own section heading ("Perspective", "Review") is strong
↪  evidence --
prefer it when it does not contradict the content.

## 3. Scope fields
```

```
- `scope_disease`: disease or pathogen
- `scope_modeling_approach`: modeling approach or method family
- `scope_social_context`: social, cultural, or institutional context
- `scope_population`: population or subgroup
- `scope_geography`: country, region, or setting
- `scope_time_period`: time period or epidemic wave
- `scope_data_source`: named data source, registry, or surveillance system

These fields record how *narrow* the paper's claims are. They are not a
↪  topic
index -- do not list everything mentioned.

The test: if the paper's central argument would no longer hold once you
↪  removed
this disease / method / population / setting, it is specific. If the
↪  argument is
general and the item merely illustrates it, it is not.

Write "NA" when the paper's discussion is general with respect to that
dimension. A review of modeling for pandemic preparedness that draws
↪  examples
from influenza, COVID-19, and Ebola is NA for disease -- the argument does
↪  not
depend on any one of them. A paper whose claims hold only for cholera in
Haiti is specific for both disease and geography.

Prefer "NA" when uncertain. Over-filling these fields is the more damaging
error, because it makes a general paper look narrow.

## 4. Quality appraisal

These four dimensions describe how well the paper is made. They do **not**
↪  affect
inclusion -- a poorly evidenced paper that meets all five criteria is
↪  still an
include, and a beautifully made one that fails a criterion is still
↪  excluded.

For each dimension below, in this order: quote the passage that most
↪  directly
supports your judgement, state your reasoning in one line, then score.

Quotes must be verbatim from the attached document, at most 40 words, and
sufficient on their own to justify the score. If no such passage exists,
↪  write
```

```
"none found" -- and score accordingly.

### Evidence Base

Are the paper's claims anchored in cited evidence, or asserted?

- **2** - Claims are consistently supported by citations to primary
↪  studies, reviews, or seminal work. Where the authors advance a
↪  contested position, they engage with the literature on both sides.
- **1** - Partly referenced. Key claims are cited but others rest on
↪  assertion or authority; or the citation set is thin, one-sided, or
↪  heavily self-referential.
- **0** - Substantially unreferenced assertion. Few or no citations
↪  supporting the central claims.

### Breadth of Discussion

Does the paper take a holistic view of modeling challenges, or treat one
↪  narrow aspect?

- **2** - Spans several distinct dimensions -- e.g. technical, data,
↪  institutional, communication, ethical -- and relates them to one
↪  another rather than listing them.
- **1** - Covers more than one aspect but stays within one register (e.g.
↪  several technical issues), or touches broader themes only in passing.
- **0** - Confined to a single narrow issue or a single aspect of
↪  modeling.

### Acknowledgment of Uncertainty & Bias

Does the paper engage seriously with uncertainty, limitations, and its own
↪  position, or advocate one-sidedly?

- **2** - Substantive treatment of uncertainty, limitations, or bias, with
↪  a balanced view that credits competing positions or acknowledges where
↪  the authors' own stance may be wrong.
- **1** - Uncertainty or limitations are acknowledged but handled briefly
↪  or formulaically; the paper leans clearly toward one position without
↪  seriously engaging alternatives.
- **0** - Little or no acknowledgment; sustained one-sided advocacy.

### Transparency

Can a reader reconstruct how the authors reached their conclusions? Apply
↪  the standard appropriate to the paper type.
```

```
- **2** - Fully transparent for its type. Review: search strategy,
↪  selection, and synthesis described. Research article: methods, data,
↪  and assumptions sufficient to replicate; code or data availability
↪  where relevant. Opinion/editorial: authors' background, role, or
↪  competing interests disclosed, and the basis for their claims made
↪  clear.
- **1** - Partially transparent. Some of the above present, material gaps
↪  remain (e.g. a review that names databases but not inclusion
↪  criteria).
- **0** - Opaque. Conclusions cannot be traced to a described process, and
↪  no relevant disclosure is offered.

## 5. Calibration

Calibration notes -- these correct the failure modes seen most often:

- Do not reward topicality. A paper being about COVID-19, pandemics, or
  preparedness says nothing about whether it meets the criteria.
- Do not reward the abstract's framing. Papers routinely promise policy
  relevance in the abstract and deliver none in the body. Score the body.
- A score of 1 is a substantive finding, not a hedge. If you cannot state
↪  what
  is present *and* what is missing, the paper is a 0 or a 2, not a 1.
- The five criteria are independent requirements. A paper routinely meets
↪  some
  and fails others; that is the normal case, not a sign you have misread
  ↪  it. Do
  not let a strong showing on one criterion soften your verdict on
  ↪  another.
- Criteria 4 and 5 are the pair most often confused. Criterion 4 asks
↪  whether
  the paper discusses limitations and challenges at all. Criterion 5 asks
  whether it is *about* something wider than its own model. A single-model
  ↪  paper
  with a thoughtful limitations section meets 4 and fails 5.
- If you cannot locate supporting text for a score of 1 or 2, the score is
↪  0.
  Absence of evidence in the document is evidence of absence for these
  ↪  purposes.
- The comment fields are for the screener to audit your reasoning. Write
  compressed noun phrases, not sentences: "policy framing in abstract
  ↪  only;
  body is parameter estimation".

## 6. Decision
```

```
- `include`   - all five criteria met.
- `exclude`   - any criterion not met.
- `uncertain` - no criterion failed, but one or more could not be settled.

Derive the decision from the verdicts you have already given; do not
↪  re-litigate
them here. If applying the rule produces a decision that feels wrong, keep
↪  the
rule and say why in `decision_rationale` -- that disagreement is a signal
↪  about
the instrument, and suppressing it hides the problem.

Return the structured object. Every field is required.
```

Supplementary Note 16: Detailed Delphi round procedures

The rating scale Both quantitative rounds used the same six-point agreement scale, presented to panellists with this stem and these anchors:

> Also rate your agreement with the statement in its current form on a scale of 1 – completely disagree, 2 – mostly disagree, 3 – somewhat disagree, 4 – somewhat agree, 5 – mostly agree, 6 – completely agree.

The scale is fully labelled and has an even number of points, so it carries no neutral midpoint: every response falls on the disagreeing side (1 to 3) or the agreeing side (4 to 6). The top-two share reported in Supplementary Note 5 is the percentage choosing 5 or 6, that is, "mostly agree" or "completely agree." Round 4, which ran in LimeSurvey, offered a seventh option, "Not qualified to respond," coded 7 and treated throughout as an abstention rather than as a rating: it is excluded from every median, interquartile range, and share, and does not count towards n. Round 3, which ran on structured text documents, had no such option, and a panellist with no view left the field blank, which is why n varies by statement in that round too.

Round 1: Internal feedback With the infoXpand consortium group ($N = 6$), namely the five consortium members who are also panellists together with A.C.V., we reviewed the initial set of 17 statements, a ranking task, and five open-ended questions. The team provided written feedback

on clarity, relevance, completeness, and redundancy. Revising against that feedback produced 26 statements, a seven-item ranking task, and seven open-ended questions for Round 2. Rounds 1 and 2 used structured text documents and a single prioritisation ranking.

Round 2: External panel feedback We distributed the 26 statements, the ranking task, and seven open-ended questions to 10 external experts identified through consortium networks. Panellists provided free-text feedback on each statement. We used only aggregated feedback to guide revisions and expanded the ranking task to three phase-specific rankings (before, during, and after a pandemic), producing 42 statements for Round 3.

Round 3: Rating and feedback Panellists ($N = 18$) rated each of the 42 statements on a six-point Likert scale, completed the three ranking tasks, and responded to seven open-ended questions. Round 3 used the same structured text documents as Rounds 1 and 2, one per panellist, returned by email. We shared a preliminary manuscript with co-authors in parallel, so some panellists rated statements while holding a draft that had already interpreted them, which is an additional anchoring risk we did not control for. Rounds 1 to 3 were identifiable, so responses could be linked across those rounds; Round 4 was anonymous, which breaks the chain and is why we report no retention figure.

Round 4: Final rating We distributed the consolidated statements for final rating to the broadest group. Authors who are not panel members received the link but did not rate it, so all analysed responses come from panel members or external experts ($N = 41$ submitted responses). Before rating, Round 4 respondents received the consolidated Round 3 feedback report. For each of the 42 Round 3 statements, it gave the mean, median, standard deviation, range, and full distribution of the Round 3 ratings, together with synthesised thematic summaries of the free-text comments rather than the comments themselves. It reported no interquartile ranges and did not classify statements against the consensus rule, but respondents could still see that most Round 3 medians were already at or near the top of the scale, which is an anchoring mechanism we did not control for. Round 4 used LimeSurvey with the same six-point scale. Unlike Round 3, free-text comments were collected once per thematic section rather than per statement. Panellists' responses

were anonymous to one another throughout. Round 4 was an anonymous online survey, so its respondents cannot be linked to earlier rounds. Free-text comments from Rounds 3 and 4 were synthesised into recurring themes with a DeepSeek-V4 model hosted locally, so no panellist-authored text left institutional control (Supplementary Note 15).

Timeline Round 1: 16 December 2025 to 6 February 2026. Round 2: 20 February to 13 March 2026. Round 3: 30 March to 20 April 2026. Round 4: 4 May to 22 May 2026. These are the stated field periods. Two Round 4 responses arrived after the survey's stated closing date, on 25 May and 5 June 2026, and we retained both in the analysis. Each round was followed by a two-week revision period. Upon completion, we circulated the final manuscript to all contributing panellists to confirm co-authorship.

Supplementary Note 17: Initial Delphi statement set (Round 1)

Supplementary Table 9 lists the 17 candidate statements we presented to the panel in Round 1, before the iterative revisions documented above. How this set was drafted and how the team rewrote it before any panellist saw it are described in Supplementary Note 15.

Supplementary Table 9: **The initial Round 1 statement set.** The 17 candidate statements derived from the systematic review and presented to the panel in Round 1, before the iterative revisions documented in the round procedures. Statements are transcribed verbatim.

#	Statement
Section A: Model design and complexity	
1	Balancing model complexity with computational speed is crucial for providing timely and reliable insights to support the rapid evolution of pandemic decision-making, even when simplifying assumptions risk omitting important epidemiological or social mechanisms.
2	Integrating socio-economic and behavioral factors into epidemiological models is crucial for capturing the real-world drivers that influence disease spread and the impacts of interventions, despite the growing demands for data.
3	Models must explicitly account for multiple pathogen strains and mutation dynamics to remain relevant and predictive across changing epidemiological contexts.
4	Agent-based models (ABM) are less suitable for short-term prediction, but valuable for understanding intervention scenarios.
Section B: Data availability and quality	
5	Timely data availability is more critical for effective modeling and decision support than completeness, even if it entails quality trade-offs that increase model uncertainty.
6	Open sharing of models, data, and assumptions, both within the scientific community and with policymakers, improves trust, reproducibility, and the practical uptake of model results.
7	Socio-behavioral data collection should be prioritized alongside clinical surveillance, despite challenges in measurement consistency and concerns about privacy.
8	Data privacy protections should be balanced against emergency data-sharing needs, with predefined ethical safeguards established prior to the pandemic.
Section C: Uncertainty, validation, and communication	
9	Uncertainties in model structure, assumptions, and data should always be explicitly quantified and communicated using standardized formats.
10	Model validation must be continuously updated with emerging empirical data, even when frequent revisions challenge policymakers' expectations of stability.
11	Sensitivity analyses should accompany all policy-informing model outputs to identify parameters that most strongly influence predictions.
12	Policymakers require structured training on interpreting model uncertainty to reduce miscommunication-driven policy errors.
13	Interactive visualization tools and dynamic dashboards should replace static reports to improve comprehension of model scenarios and limitations.
Section D: Collaboration, interdisciplinarity, and ethics	
14	Effective pandemic preparedness necessitates long-term, interdisciplinary collaboration among modellers, social scientists, economists, behavioral experts, and public health practitioners.
15	Dedicated translator roles should serve as a mediator between scientific modeling teams and policymakers, reducing misinterpretation and increasing policy uptake.
16	Ethical guidelines must be explicitly incorporated into model development and scenario communication, particularly regarding resource allocation and the impacts on vulnerable groups.
17	Global modeling cooperation should be strengthened through joint platforms, funding mechanisms, and cross-border data-sharing agreements.

Supplementary Note 18: Thematic synthesis

Order of operations The statement set was drafted and rewritten in full before this synthesis began (Supplementary Note 15). The eight themes were later derived from the same extracted items and were not used as input for statement drafting. Any correspondence between themes and statements is, therefore, retrospective, and Table 2 of the main article presents it as such.

Procedure The synthesis used the recommendation and challenge fields extracted from each of the 159 publications in the initial screen and stored in the review summary workbook. Splitting those fields into individual items gives the 568 recommendation items and 583 challenge items reported in the results. We grouped items into candidate themes with large-language-model assistance (Anthropic Claude Opus 4.6), merged overlapping candidates, and, as a team, reviewed and accepted the resulting set, arriving at eight themes.

We later re-derived theme coverage across the 135 publications in the review reports because the original synthesis predated the eligibility reassessment, and its per-publication assignments were never recorded. Coverage is now determined by a stated term list applied to each publication's extracted challenge and recommendation text (Supplementary Table 10), which is rerunnable and can be checked case by case.

We developed and validated that list on two separate hand-labelled samples, in that order. One author (L. Stellbrink) labelled a first random sample of 30 publications, and we revised the term list against its errors in two passes, correcting terms that fired on the wrong theme and adding vocabulary that the corpus used but the list had missed. Because a list tuned on a sample cannot be scored on it, the same author then labelled a second random sample of 30, and we report performance on that one. Against those held-out labels, the term list achieves precision of 0.82, recall of 0.62, and Cohen's $\kappa = 0.55$. Five of the second 30 had also appeared in the first, so they are not strictly held out. Excluding them yields $\kappa = 0.53$. Both samples were labelled by one author, so κ measures agreement between the codebook and a single-coder reference standard, not between two independent coders. A second coder on a subsample would establish whether the reference standard is itself reproducible, and we did not do this; it is a limitation of the theme shares rather than of the consensus results, which do not depend on them. Recall below precision means the reported shares are conservative, since publications are missed more often than wrongly assigned.

The eight themes themselves are unchanged from the original synthesis and should be read as a single defensible organisation of the reviewed material rather than a partition that another team would necessarily reproduce. Nothing in the consensus results depends on them: we drafted, revised, and rated the statements independently of this step.

Supplementary Table 10: **Codebook for thematic coverage.** A publication counts as addressing a theme where its extracted challenge and recommendation text matches at least one term listed for that theme. Terms match case-insensitively and on word stems, so `data shar` covers sharing and shared. An ellipsis marks a gap within the same sentence, bounded per term at between 20 and 40 characters, and a slash separates alternatives. The exact expressions are in the deposited analysis script.

Theme	**Terms**
Data governance, sharing, and infrastructure	*governance*: `data govern`, `privacy`, `data protection`, `data use agreement`, `consent`, `ethical ... data` *sharing*: `data shar`, `data access`, `data availab`, `open data`, `FAIR`, `data linkage`, `linked data` *infrastructure*: `data infrastructure`, `surveillance system / platform / network`, `real-time data`, `interoperab`, `data pipeline`, `data registry / repositor`, `routine data`, `health data`, `data collection`, `data quality`, `data standard`
Uncertainty quantification and communication	*quantification*: `uncertaint`, `confidence interval`, `credible interval`, `probabilistic`, `prediction interval`, `structural uncertainty`, `parameter uncertainty`, `biases ... estimat / inference` *communication of uncertainty*: `communicat ... uncertaint`, `convey ... uncertaint`, `uncertaint ... communicat`
Public and stakeholder communication	*public*: `public communicat`, `communicat ... public`, `media`, `public trust`, `public understanding`, `misinformation`, `lay audience / public / reader`, `general public`, `public perception` *stakeholder*: `stakeholder`, `end-user ... engag / communicat / need`, `engag ... public / communit / stakeholder`, `co-produc`, `co-develop` *presentation*: `visualisation / visualization`, `plain language`, `interactive tool / dashboard`, `communicat ... non-technical / non-expert`

continued on next page

Supplementary Table 10: *(continued)*

Theme	Terms
Science-policy interface	*actors*: policy maker, decision maker, health agenc, public health authorit, end-user *interface*: science-policy, advisory, decision support, policy impact, policy process, translat ... polic, operational need, inform ... decision / polic, actionab
Model development and methodological improvement	*model choice*: model complexity, model selection, hybrid model, agent-based, compartmental, individual-based, machine learning, fit for purpose, choice of model, homogeneous mixing, oversimplified assumption, structural model *methodological improvement*: calibrat, parametris / parametriz, parameter estimation, model fitting, MCMC, identifiabilit, sensitivity analys, stochastic, model criticism, ensemble, multi-model, model validation, model comparison, methodolog ... improve / advance / develop
Capacity building and global equity	*capacity building*: capacity ... build / develop / strengthen / constrain, local capacity, workforce, training (but not training data, training set), curricul, fellowship, sustained funding, skills *global equity*: low- and middle-income, LMIC, global south / global equity, equity, inequalit, resource-limited / constrained / poor, under-resourced, north-south, vulnerable population, developing countr, Africa, underrepresent, geographic ... bias / coverage / representat / disparit
Model transparency and reproducibility	*transparency*: transparen, black box, decipher model, assumptions ... explicit / stated / document / clear, opaque *reproducibility*: reproducib ... model, model ... reproducib, replicat ... model / result / analys, open science, open source, model code, code and/or documentation, publish ... code, open code, version control, auditab, re-run

continued on next page

Supplementary Table 10: *(continued)*

Theme	Terms
Reporting standards and documentation	*reporting standards*: `reporting standard / guideline / item / checklist / practice`, `checklist`, `EPIFORGE`, `CHEERS`, `minimum reporting`, `standardised / standardized reporting`, `ODD protocol`, `reporting ... consisten / complete` *documentation*: `documentation`, `document ... method / assumption / parameter / data source / model`, `metadata`, `provenance`, `model description`

Supplementary Note 19: Reproducibility and analysis pipeline

We implemented the Round 4 analysis as a documented, rerunnable pipeline and deposited it (https://doi.org/10.5281/zenodo.22679573) alongside the anonymised survey data (https://doi.org/10.5281/zenodo.22232256), so that anyone can regenerate the statistics, tables, and figures in this paper from those inputs. One category of data is deliberately withheld. The free-text comments panellists wrote in Rounds 3 and 4 are not deposited and remain under institutional control, because we collected them under a confidentiality commitment and they can carry identifying detail. They reach this article only as the synthesised themes reported in Supplementary Note 7, and the Data availability statement is scoped to the quantitative data for that reason. Everything the reported results depend on is deposited. Two data files are required: the unmodified survey export and a single reconstructed record for the one panellist whose final-round submission was split across partial attempts after a technical interruption. The reconstruction matches the attempts to self-reported demographics and the panellist's own indication of completed statements, and a build step deterministically removes superseded rows and abandoned reattempts, identified by exact demographic matches and overlapping start times. The script lists every removed response identifier with its reason, and running it on those two files reproduces the analysis dataset exactly.

Response handling and consensus rules A Round 4 response counted as complete if the survey recorded a submission date for it; blank optional

answers did not affect that. Responses with at least one rating were retained for the statements they rated, whether or not they were submitted, so the number of raters varies by statement. We treated "not qualified to respond" as an abstention and excluded it from all statistics. Where a panellist submitted more than once, we consolidated into a single response, a documented deviation from the rule pre-registered at osf.io/p7wm3. Because Round 4 was distributed as an anonymous, shareable link, that consolidation relies on exact demographic matches and overlapping start times, which cannot distinguish between one person submitting twice and two panellists who share a demographic profile and started at the same time. We judged the risk small, given the granularity of the demographic fields and the overlap in timing, but it is not eliminated.

We assessed stability between Rounds 3 and 4 descriptively, treating a statement as stable where the median moved by less than 1 and the IQR by less than 0.5, and we report the outcome here rather than using it to classify. Thirteen of the 40 statements meet both parts. Fifteen fail the median part and 25 the IQR part, and B07 moves furthest on each (median: 3.5 to 6; IQR: 4.0 to 1.0). The criterion is therefore not met for most of the set. The design predicts this, and it does not signal unreliable ratings: the statements were reworded between rounds, two pairs were merged, and Round 4 drew on a broader, partly different panel, so the two rounds are not paired measurements of the same items by the same people. The direction of the movement is uniform, which a chance process would not produce: every statement either held its consensus class or moved towards consensus, none moved away, no median fell, and B05 alone shows any rise in dispersion (IQR 0 to 1). We therefore read the comparison descriptively, as Supplementary Note 4 does, and rest the consensus classification on Round 4 alone.

The pipeline extracts per-statement ratings, computes medians, interquartile ranges, and consensus classifications, compares rounds, aggregates rankings, and regenerates all tables and figures, writing machine-readable outputs at each stage. One Round 3 response required a manual correction: for D07, on the well-being of children and other vulnerable groups, a rating misplaced in the free-text field was restored, taking that statement's Round 3 n from 16 to 17 and its IQR from 1.25 to 2. Under the three bands in the caption of Supplementary Table 4, this moves the statement's Round 3 reading from near-consensus to non-consensus. It leaves the Round 4 classification on which every reported outcome rests untouched, and D07 reaches consensus at Round 4 either way.

Reconciling the Round 4 denominators Several counts circulate in the main article because they answer different questions, and they are consistent: each is a subset of the one before it. Round 4 produced 68 raw responses. Six were removed: three fragments of one interrupted submission and three re-attempts abandoned before submission. Consolidating those three fragments into a single reconstructed record adds one back, so 68 minus 6 plus 1 leaves 63 analysed responses. Of those, 51 rated at least one statement, 47 answered all 40 items, and 41 submitted the survey and are therefore complete; these are reported as the panel size N.

Answering all 40 items is not the same as being counted in all 40 denominators, because "Not qualified to respond" is an answer but not a rating. Thirty-six abstentions fell across 19 of the 40 statements: five on C04; four on A11; three on D05; two each on A07, A08, A09, B08, C03, C07, D06, and D08; and one each on A03, A04, A05, A06, A10, B06, C01, and C02. Each statement was answered by between 47 and 51 panellists, while the per-statement denominators n run from 43 to 51 once abstentions are removed, and the ranking phases from 45 to 46. This is why a statement can carry an n below 47 even though 47 people answered every item.

Answering every item is likewise not the same as completing the survey: six panellists answered all 40 without submitting, which is why 47 answered the full set while only 41 are complete. We define the panel by submission because submission is the point at which a panellist confirmed the response as theirs. The completers-only reanalysis in Supplementary Note 5 shows that restricting every statistic to those 41 changes no consensus classification. A reader who prefers the 47 can read that reanalysis as the bound in the other direction.

Supplementary Note 20: PRISMA 2020 checklist

Supplementary Table 11 reports this review against the 27 items of the PRISMA 2020 statement.[21] The item wording is reproduced verbatim, so it keeps the spelling of the original; only the location column is ours.

Where an item cannot apply to a review of this kind or was not done, the entry says so and why, rather than sitting blank. Three groups fall outside this review. The items that presuppose extracted outcome data or effect estimates have nothing to report (10a, 12, 19, 20d), because the review extracted the challenges and recommendations each publication discusses, not results. For the same reason, there was no statistical synthesis to describe

(20b, and the meta-analysis half of 13d). No risk-of-bias, reporting-bias, or certainty appraisal was made (11, 14, 15, 18, 21, 22); Supplementary Note 12 sets out the relevance framework used in place of the first of these and why no validated tool fits a corpus running from commentaries to research articles. Three items, finally, are not met, rather than inapplicable: the title does not identify the report as a systematic review (item 1), and the review was neither registered nor governed by a written protocol (items 24a and 24b). Methods, in the "Systematic review" of the main article, state the last of these as one of three deliberate adaptations.

Reporting the review against the checklist after the fact does not make it a protocol-driven review, nor is the table offered as one. It records where each item is addressed and says plainly where it is not.

Supplementary Table 11: **PRISMA 2020 checklist for the systematic review.** The 27 items of the PRISMA 2020 statement [21] against the review reported here. The checklist item wording is reproduced verbatim and so keeps the spelling of the original; only the location column is ours. Section names refer to the main article unless a Supplementary Note or Table is named. Items that a review of this kind cannot report, and the three that were not done, say so and why rather than being left blank (Supplementary Note 20).

Section and topic	Item	Checklist item	Location where item is reported
Title			
Title	1	Identify the report as a systematic review.	Not done. The title names the Delphi consensus, which is the study's principal output. The abstract and the opening of the Main text both state that a systematic review supplied the material the panel rated.
Abstract			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	Abstract. Nature Health requires an unstructured, unreferenced abstract of at most 150 words, so the PRISMA for Abstracts items are not separately signposted. The abstract gives the design, the panel size, the principal findings, and the limitation of panel composition.
Introduction			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	Main, paragraphs 1 to 3.
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	Main, paragraph 4.
Methods			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	Methods, "Systematic review" and "Screening and selection"; Supplementary Note 13 gives the five criteria with the operational test applied to each. Publications were not grouped for separate syntheses; the retrospective thematic grouping is in Methods, "Thematic synthesis", and Supplementary Note 18.
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	Methods, "Systematic review": PubMed, Scopus, and Web of Science, each last searched on 7 February 2025. No registers, websites, organisations, or reference lists were consulted, and no update search was run (Supplementary Note 14).

continued on next page

Supplementary Table 11: *(continued)*

Section and topic	Item	Checklist item	Location where item is reported
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	Supplementary Note 14 reproduces all three queries verbatim, with the subject-classification filter each carries, the English-language restriction applied at screening, and two stated limitations of the query set.
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	Methods, "Screening and selection". One author screened every title and abstract, so that stage was not independently duplicated. Full texts were assessed with a structured LLM-assisted protocol treated as a first pass, with the authors deciding; one author verified 84 of the 159 reassessed records, including every proposed exclusion. Supplementary Note 15 gives the models, the prompt, and the limits of that verification.
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	Methods, "Data extraction", and Supplementary Note 15. One LLM-assisted extraction pass per record, reviewed by the authors and not duplicated by a second extractor. Study investigators were not contacted.
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	Not applicable. The review sought no outcome data. It extracted the challenges and recommendations each publication discusses (Methods, "Data extraction").
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	Methods, "Data extraction", and Supplementary Note 15 list the extracted fields; Methods, "Relevance framework", and Supplementary Table 8 give the four relevance dimensions and their anchors. Assumptions about missing or unclear information were not separately recorded.

continued on next page

Supplementary Table 11: *(continued)*

Section and topic	Item	Checklist item	Location where item is reported
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	Not assessed. No validated tool fits a corpus running from commentaries to research articles, so relevance was appraised in its place (Methods, "Relevance framework"; Supplementary Note 12 and Supplementary Table 8). Relevance was scored once per publication with LLM assistance, is not a risk-of-bias judgement, and the framework itself is unvalidated, which the Discussion states as a limitation.
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	Not applicable. The review synthesised text, not estimates, so there are no effect measures.
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	Methods, "Thematic synthesis". Every included publication entered one thematic synthesis, so there were no separate syntheses to assign publications to.
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	Methods, "Thematic synthesis", and Supplementary Note 18: the extracted challenge and recommendation fields were split into individual items, and theme coverage was re-derived from a stated term list (Supplementary Table 10).
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	Results, "Evidence base from the systematic review" and "Eight themes in the reviewed literature"; Supplementary Tables 1, 2, and 3.
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	Methods, "Thematic synthesis", and Supplementary Note 18 describe the grouping of extracted items into eight themes, the LLM assistance used, and the term-based coverage codebook with its validation. No meta-analysis was performed, the material being text rather than estimates.
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	Not applicable to study results, there being no effect estimates. Theme coverage is compared between publications before and from 2020 onwards; Results, "Evidence base from the systematic review".

continued on next page

Supplementary Table 11: *(continued)*

Section and topic	Item	Checklist item	Location where item is reported
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	No synthesised statistical result to test. Two robustness checks on the review are reported instead: the sharpened eligibility criteria reapplied to all 159 publications (Methods, "Screening and selection", and Fig. 2 of the main article), and the codebook agreement recomputed with the overlapping hand labels removed (Supplementary Note 18).
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	Not assessed. Reporting bias in this sense presupposes a synthesis of results, which this review does not have. The English-language restriction, the single search date, and the known gaps in the queries are stated in Methods, "Systematic review", Supplementary Note 14, and the Discussion.
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	Not assessed. No GRADE or equivalent appraisal was made. The review supplied candidate statements, and the strength of expert agreement on them is what the study reports instead (Methods, "Rating and response handling" and "Consensus criterion").
Results			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	Results, "Evidence base from the systematic review"; Fig. 2 of the main article (PRISMA 2020 flow diagram); Methods, "Screening and selection".
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	Fig. 2 of the main article gives the number excluded at full text with the criterion each failed, for the initial assessment (107) and the reassessment (24). Excluded publications are not cited individually; the screening export naming every record and its decision is deposited (Data availability).
Study characteristics	17	Cite each included study and present its characteristics.	Supplementary Note 1 and Supplementary Table 1 give period, type, disease focus, and relevance band; Supplementary Note 2 and Supplementary Table 2 cite all 135 included publications with type and themes.

continued on next page

Supplementary Table 11: *(continued)*

Section and topic	Item	Checklist item	Location where item is reported
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	Not assessed; see item 11. The distribution of relevance bands across the 135 publications is in Supplementary Table 1.
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	Not applicable; no outcome data or effect estimates were extracted. What was extracted per publication, its theme coverage, appears in Supplementary Table 2.
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	Results, "Evidence base from the systematic review" and "Eight themes in the reviewed literature"; Supplementary Note 1 and Supplementary Table 1 describe the contributing publications. No risk-of-bias assessment was made (item 11).
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	No statistical synthesis was conducted. Theme coverage across the 135 publications is reported in Results, "Eight themes in the reviewed literature", Supplementary Note 3, and Supplementary Table 3.
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	Results, "Evidence base from the systematic review" reports theme coverage before and from 2020 onwards; see item 13e.
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	Not applicable; see item 13f for the two robustness checks the review does report.
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	Not assessed; see item 14.
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	Not assessed; see item 15.
Discussion			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	Discussion, which sets the recommendations beside the Lancet Commission, the EPIFORGE reporting checklist, and the WHO preparedness framework, and which reports how long the same concerns have been raised.

continued on next page

Supplementary Table 11: *(continued)*

Section and topic	Item	Checklist item	Location where item is reported
	23b	Discuss any limitations of the evidence included in the review.	Discussion, limitations: COVID-19 dominates the corpus, only English-language publications were included, the relevance framework is unvalidated, and the theme codebook agrees only moderately with hand labels. Supplementary Notes 12 and 18.
	23c	Discuss any limitations of the review processes used.	Discussion, limitations, and Methods, "Systematic review", which states the three adaptations: single-screener title and abstract screening, relevance appraisal in place of risk of bias, and no registered review protocol. Supplementary Notes 14 and 15.
	23d	Discuss implications of the results for practice, policy, and future research.	Discussion; Results, "Ten priorities for the inter-pandemic period"; Table 3 of the main article.
Other information			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	The review was not registered. The Delphi analysis plan was pre-registered at OSF (osf.io/p7wm3), which is not a review registration; Methods, "Systematic review".
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	A review protocol was not prepared; Methods, "Systematic review".
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	Not applicable, there being no registration and no protocol. The one substantive change of course is reported anyway: criteria 4 and 5 were given explicit tests and all 159 publications were rescreened (Methods, "Screening and selection"; Supplementary Note 13; Fig. 2 of the main article).
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	Funding, which also states that the funders had no role in the study.
Competing interests	26	Declare any competing interests of review authors.	Competing interests.

continued on next page

Supplementary Table 11: *(continued)*

Section and topic	Item	Checklist item	Location where item is reported
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	Data availability and Code availability. Deposited at Zenodo: the full-text screening export underlying the review and the hand labels used to validate the codebook (10.5281/zenodo.22232256); the screening instrument holding the five criteria and the decision rule, which is the template applied to every record (10.5281/zenodo.22647106); and the analysis code that builds every computed table and figure (10.5281/zenodo.22679573).

Supplementary Note 21: Dissemination and keeping the set current

We intend to disseminate these recommendations broadly to the communities they serve. The primary output is this peer-reviewed publication, co-authored by contributing panellists. To maximise accessibility and impact, we plan the following dissemination activities:

- **Open-access publication** to ensure unrestricted availability to researchers, practitioners, and policymakers worldwide.
- **Plain-language summary** accompanying the publication, suitable for communication to non-specialist audiences, including policymakers and media.
- **Targeted dissemination** to public health agencies, national pandemic preparedness bodies, and international organisations, including the World Health Organization and the European Centre for Disease Prevention and Control.
- **Open data and materials**, including the anonymised survey data, the term-based theme codebook of Supplementary Note 18 (Supplementary Table 10), the relevance framework of Supplementary Note 12 (Supplementary Table 8), and supplementary analyses, deposited in a public repository.

The consensus set is intended to be revised as evidence accumulates rather than fixed as a static checklist. We recommend revisiting it periodically as new evidence emerges, incorporating lessons from future outbreaks and advances in modelling methodology.

References

1. von der Gracht, H. A. Consensus measurement in Delphi studies: Review and implications for future quality assurance. *Technological Forecasting and Social Change* **79**, 1525–1536 (2012).

2. Beiderbeck, D., Frevel, N., von der Gracht, H. A., Schmidt, S. L. & Schweitzer, V. M. Preparing, conducting, and analyzing Delphi surveys: Cross-disciplinary practices, new directions, and advancements. *MethodsX* **8**, 101401 (2021).

3. Niederberger, M., Köberich, S. & members of the DeWiss Network. Coming to consensus: the Delphi technique. *European Journal of Cardiovascular Nursing* **20**, 692–695 (2021).

4. Jünger, S., Payne, S. A., Brine, J., Radbruch, L. & Brearley, S. G. Guidance on Conducting and REporting DElphi Studies (CREDES) in palliative care: Recommendations based on a methodological systematic review. *Palliative Medicine* **31**, 684–706 (2017).

5. Majeed, A., Maile, E. J. & Bindman, A. B. The primary care response to COVID-19 in England's National Health Service. *Journal of the Royal Society of Medicine* **113**, 208–210 (2020).

6. Starfield, B., Shi, L. & Macinko, J. Contribution of Primary Care to Health Systems and Health. *The Milbank Quarterly* **83**, 457–502 (2005).

7. Van Weel, C. & Kidd, M. R. Why strengthening primary health care is essential to achieving universal health coverage. *Canadian Medical Association Journal* **190**, E463–E466 (2018).

8. Páscoa, R., Rodrigues, A. P., Silva, S., Nunes, B. & Martins, C. Comparison between influenza coded primary care consultations and national influenza incidence obtained by the General Practitioners Sentinel Network in Portugal from 2012 to 2017. *PLOS ONE* **13**, e0192681 (2018).

9. Grimm, V. *et al.* A standard protocol for describing individual-based and agent-based models. *Ecological Modelling* **198**, 115–126 (2006).

10. Willem, L., Verelst, F., Bilcke, J., Hens, N. & Beutels, P. Lessons from a decade of individual-based models for infectious disease transmission: a systematic review (2006-2015). *BMC Infectious Diseases* **17**, 612 (2017).

11. Sauerbrei, W. *et al.* STRengthening analytical thinking for observational studies: the STRATOS initiative. *Statistics in Medicine* **33**, 5413–5432 (2014).

12. Rivers, C. *et al.* Using "outbreak science" to strengthen the use of models during epidemics. *Nature Communications* **10**, 3102 (2019).

13. Shea, B. J. *et al.* AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ* **358**, j4008 (2017).

14. Munn, Z., Stern, C., Aromataris, E., Lockwood, C. & Jordan, Z. What kind of systematic review should I conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences. *BMC Medical Research Methodology* **18**, 5 (2018).

15. von Elm, E. *et al.* The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *The Lancet* **370**, 1453–1457 (2007).

16. Zhang, T. *et al.* Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics* **12**, 39–57 (2024).

17. Croxford, E. *et al.* Evaluating clinical AI summaries with large language models as judges. *npj Digital Medicine* **8**, 640 (2025).

18. Gallifant, J. *et al.* The TRIPOD-LLM reporting guideline for studies using large language models. *Nature Medicine* **31**, 60–69 (2025).

19. Huang, L. *et al.* A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems* **43**, 42:1–42:55 (2025).

20. Asgari, E. *et al.* A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *npj Digital Medicine* **8**, 274 (2025).

21. Page, M. J. *et al.* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* **372**, n71 (2021).