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

Leonard Stellbrink

leonard.stellbrink@uni-luebeck.de

University of Luebeck https://orcid.org/0000-0003-2188-1743

Hendrik Nunner

University of Luebeck

Anne Schönberger

University of Luebeck

Maged Mortaga

University of Luebeck

Michael Mäs

Karlsruhe Institute of Technology

Kai Nagel

Technical University of Berlin

Sydney Paltra

Technical University of Berlin

Ben Ashby

Simon Fraser University

Philippe Beutels

Antwerpen

Tom Britton

SU

Sarah Cuschieri

University of Malta

Thomas Czypionka

Institute for Advanced Studies

Uga Dumpis

University of Latvia

Elizabeth Fearon

University College London

Neil M. Ferguson

Imperial College London https://orcid.org/0000-0002-1154-8093
Fakhteh Ghanbarnejad
SRH University of Applied Sciences Heidelberg
Enrico Glaab
University of Luxembourg
Reanna Gregory
University of Warwick
Eva Grill
Ludwig-Maximilians-University Munich
Claudia Hanson
Karolinska Institute https://orcid.org/0000-0001-8066-7873
Niel Hens
UHasselt https://orcid.org/0000-0003-1881-0637
Thomas House
Manchester
Lorenzo Pellis
University of Manchester
Mark Jit
NYU
André Karch
University of Muenster
Peter Klimek
Medical University of Vienna https://orcid.org/0000-0003-1187-6713
Tyll Krüger
Wrocław University of Science and Technology
Alexander Kuhlmann
University of Luebeck
Martin Kühn
German Aerospace Center, University of Bonn
Jana Lasser
University of Graz
Nicola Low
UniBe
Carlos Martins
University of Porto
James McCaw
The University of Melbourne
Martin McKee
LSHTM

Rafael Mikolajczyk
Martin Luther University Halle-Wittenberg
Christina Pagel
University College London
Jasmina Panovska-Griffiths
Pandemic Sciences Institute, Nuffield Department of Medicine, University of Oxford
https://orcid.org/0000-0002-7720-1121
Daniela Paolotti
ISI https://orcid.org/0000-0003-1356-3470
Matjaž Perc
University of Maribor
Elena Petelos
University of Crete
Martyn Pickersgill
University of Edinburgh
Michael Plank
Canterbury NZ https://orcid.org/0000-0002-7539-3465
Niki Popper
Technical Universtity Vienna
Barbara Prainsack
University of Vienna
Joacim Rocklov
Heidelberg University https://orcid.org/0000-0003-4030-0449
Francesca Scarabel
University of Leeds
Helena Stage
University of Bristol
Anthony Staines
Dublin City University
Ben Swallow
University of St Andrews
Ewa Szczurek
University of Warsaw
Robin Thompson
Mathematical Institute, University of Oxford https://orcid.org/0000-0001-8545-5212
Michael Tildesley
Warwick
Daniel Villela
Programa de Computação Científica

Mirjam Kretzschmar
UMC Utrecht
Viola Priesemann
University of Goettingen
André Calero Valdez
University of Luebeck

Article

Keywords:

Posted Date: September 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10993323/v1

Additional Declarations: Yes there is potential Competing Interest. The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper. The following authors declare the advisory and committee roles listed below as interests: T. House and E. Fearon are members of the Scientific Pandemic Infections Group on Modelling (SPI-M), a committee that advises the UK Department of Health and Social Care. J. M. McCaw is an invited expert member of the Communicable Diseases Network Australia, the Australian government's committee tasked with providing national public health coordination and leadership. M. McKee and C. Pagel are members of Independent SAGE, a group of volunteers that conducted public engagement during the pandemic. M. Pickersgill is a member of the Scottish Science Advisory Council. This article was written in an independent capacity.

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

Leonard Stellbrink[1,*], Hendrik Nunner[1], Anne Schönberger[1], Maged Mortaga[1], Mirjam E. Kretzschmar[2,3,4], Michael Mäs[5], Kai Nagel[6], Sydney Paltra[6], Viola Priesemann[7], Ben Ashby[8], Philippe Beutels[9], Tom Britton[10], Sarah Cuschieri[11], Thomas Czypionka[12], Uga Dumpis[13], Elizabeth Fearon[14,15], Neil M. Ferguson[16], Fakhteh Ghanbarnejad[17], Enrico Glaab[18], Reanna Gregory[19,20], Eva Grill[21], Claudia Hanson[22], Niel Hens[23,9], Thomas House[24], Mark Jit[25], André Karch[26,4], Peter Klimek[27,28,29], Tyll Krüger[30], Alexander Kuhlmann[31], Martin Kühn[32,33], Jana Lasser[34], Nicola Low[35], Carlos Martins[36], James M. McCaw[37], Martin McKee[38], Rafael Mikolajczyk[39,40], Christina Pagel[41], Jasmina Panovska-Griffiths[42,43], Daniela Paolotti[44], Lorenzo Pellis[24], Matjaž Perc[45,46,47,48], Elena Petelos[49,50], Martyn Pickersgill[51], Michael J. Plank[52], Niki Popper[53], Barbara Prainsack[54], Joacim Rocklöv[55], Francesca Scarabel[56], Helena Stage[57], Anthony Staines[58,†], Ben Swallow[59], Ewa Szczurek[60], Robin N. Thompson[61], Mike Tildesley[20], Daniel A. M. Villela[62,4], André Calero Valdez[1,*]

[1] Institute of Human-Centered Interactive Systems, University of Luebeck, Ratzeburger Allee 160, Luebeck, 23562, Schleswig-Holstein, Germany

[2] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Heidelberglaan 100, Utrecht, 3584 CX, Netherlands

[3] Center for Complex Systems Studies, Utrecht University, Utrecht, Netherlands

[4] Interdisciplinary Center for Mathematical Modeling of Infectious Disease Dynamics (IMMIDD), University of Münster, Münster, Germany

[5] Institute for Sociology and Computational Social Science, Karlsruhe Institute of Technology (KIT), Kaiserstraße 12, Karlsruhe, 76131, Germany

[6] Transport Systems Planning and Transport Telematics, Technical University of Berlin, Straße des 17. Juni 135, Berlin, 10623, Germany

[7] University of Göttingen, Platz der Göttinger Sieben 1, Göttingen, 37073, Germany

[8] Department of Mathematics, Simon Fraser University, 8888 University Drive, Burnaby, V5A 1S6, BC, Canada

[9] Centre for Health Economics Research and Modelling Infectious Diseases (CHERMID), Vaccine & Infectious Disease Institute, University of Antwerp, Antwerp, 2000, Belgium

[10] Department of Mathematics, Stockholm University, Universitetsvägen 10A, Stockholm, SE-106 91, Sweden

[11] Faculty of Medicine & Surgery, University of Malta, Msida, MSD2080, Malta

[12] Institute for Advanced Studies (IHS), Josefstädter Straße 39, Vienna, 1080, Austria

[13] Faculty of Medicine and Life Sciences, University of Latvia, Riga, LV-1586, Latvia

[14] Institute for Global Health, University College London, Gower Street, London, WC1E 6BT, United Kingdom

[15] The National Institute for Health and Care Research Health Protection Research Unit in Blood Borne and Sexually Transmitted Infections at University College London in Partnership with the UK Health Security Agency, Gower Street, London, WC1E 6BT, United Kingdom

[16] Jameel Institute, School of Public Health, Imperial College London, White City Campus, 90 Wood Lane, London, W12 0BZ, United Kingdom

[17] School of Technology and Architecture, SRH University of Applied Sciences Heidelberg, Leipzig Campus, Prager Straße 40, Leipzig, 04317, Germany

[18] Biomedical Data Science Group, Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6 Avenue du Swing, Belvaux, L-4367, Luxembourg

[19] Warwick Medical School, University of Warwick, Coventry, CV4 7AL, United Kingdom

[20] Zeeman Institute for Systems Biology and Infectious Disease Epidemiology Research (SBIDER), University of Warwick, Coventry, CV4 7AL, United Kingdom

[21] Institute for Medical Information Processing, Biometry and Epidemiology (IBE), LMU Medizin, Ludwig-Maximilians-Universität München, Marchioninistr. 15, München, 81377, Germany

[22] Department of Global Public Health, Karolinska Institutet, Nobels väg 5, Stockholm, 171 77, Sweden

[23] Data Science Institute, Hasselt University, Martelarenlaan 42, Hasselt, 3500, Belgium

[24] Department of Mathematics, University of Manchester, Alan Turing Building, Oxford Road, Manchester, M13 9PL, United Kingdom

[25] Department of Global and Environmental Health, School of Global Public Health, New York University, 708 Broadway, New York, NY, 10003, USA

[26] Institute of Epidemiology and Social Medicine, University of Münster, Münster, Germany

[27] Institute of the Science of Complex Systems, Center for Medical Data Science, Medical University of Vienna, Spitalgasse 23, Vienna, 1090, Austria

[28] Complexity Science Hub, Metternichgasse 8, Vienna, 1030, Austria

[29] Supply Chain Intelligence Institute Austria, Metternichgasse 8, Vienna, 1030, Austria

[30] Wrocław University of Science and Technology, Wybrzeże Wyspiańskiego 27, Wrocław, 50-370, Poland

[31] Institute of Social Medicine and Epidemiology, University of Luebeck, Ratzeburger Allee 160, Luebeck, 23562, Germany

[32] Institute of Software Technology, Department of High-Performance Computing, German Aerospace Center (DLR), Linder Höhe, Cologne, 51147, Germany

[33] Bonn Center for Mathematical Life Sciences and Life and Medical Sciences Institute, University of Bonn, Endenicher Allee 64, Bonn, 53115, Germany

[34] IDea_Lab, University of Graz, Leechgasse 34, Graz, 8010, Austria

[35] Institute of Social and Preventive Medicine (ISPM), University of Bern, Mittelstrasse 43, Bern, 3012, Switzerland

[36] RISE-Health, Faculty of Medicine, University of Porto, Alameda Prof. Hernâni Monteiro, Porto, 4200-319, Portugal

[37] School of Mathematics and Statistics, and Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Grattan Street, Parkville, VIC, 3010, Australia

[38] London School of Hygiene & Tropical Medicine, 15-17 Tavistock Place, London, WC1H 9SH, United Kingdom

[39] Institute for Medical Epidemiology, Biometrics and Informatics, Interdisciplinary Center for Health Sciences, Medical School of the Martin Luther University Halle-Wittenberg, Magdeburger Straße 8, Halle (Saale), 06112, Germany

[40] Middle German Center for Biomedical Modelling and Simulation in Life Sciences, Halle (Saale), Germany

[41] Clinical Operational Research Unit, University College London, Taviton Street, London, WC1H 0BT, United Kingdom

[42] Pandemic Sciences Institute, University of Oxford, Oxford, OX3 7LF, United Kingdom

[43] The Queen's College, University of Oxford, Oxford, OX1 4AW, United Kingdom

[44] ISI Foundation, Via della Rocca 20, Turin, 10123, Italy

[45] Faculty of Natural Sciences and Mathematics, University of Maribor, Koroška cesta 160, Maribor, 2000, Slovenia

[46] Community Healthcare Center Dr. Adolf Drolc Maribor, Ulica talcev 9, Maribor, 2000, Slovenia

[47] Department of Physics, Kyung Hee University, 26 Kyungheedae-ro, Seoul, 02447, Republic of Korea

[48] University College, Korea University, 145 Anam-ro, Seoul, 02841, Republic of Korea

[49] Clinic of Social and Family Medicine, Faculty of Medicine, University of Crete, Heraklion, 71003, Greece

[50] Health Services Research, CAPHRI Care and Public Health Research Institute, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, Netherlands

[51] Centre for Biomedicine, Self and Society, School of Population Health Sciences,

University of Edinburgh, 1–7 Little France Road, Usher Building, BioQuarter, Edinburgh, EH16 4UX, United Kingdom

[52] School of Mathematics and Statistics, University of Canterbury, Science Road, Ilam, Christchurch, 8140, New Zealand

[53] TU Wien, Research Group Data Science & Research Group Computational Statistics, Favoritenstraße 9-11, Vienna, 1040, Austria

[54] Department of Political Science, University of Vienna, Universitätsstraße 7, Vienna, 1010, Austria

[55] Heidelberg Institute of Global Health & Institute of Scientific Computing, Heidelberg University, Im Neuenheimer Feld 205, Heidelberg, 69120, Germany

[56] School of Mathematics, University of Leeds, Leeds, LS2 9JT, United Kingdom

[57] School of Engineering Mathematics and Technology, University of Bristol, Bristol, BS8 1TW, United Kingdom

[58] School of Nursing, Psychotherapy and Community Health, Dublin City University, Dublin, D09 V209, Ireland

[59] School of Mathematics and Statistics, University of St Andrews, St Andrews, KY16 9SS, United Kingdom

[60] Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, 02-097, Poland

[61] Mathematical Institute, University of Oxford, Radcliffe Observatory Quarter, Woodstock Road, Oxford, OX2 6GG, United Kingdom

[62] Programa de Computação Científica, Fundação Oswaldo Cruz, Av. Brasil 4365, Rio de Janeiro, 21040-900, Brazil

*Corresponding authors: Leonard Stellbrink, André Calero Valdez
leonard.stellbrink@uni-luebeck.de, andre.calerovaldez@uni-luebeck.de

†Deceased: Anthony Staines.

The Delphi expert panel that provided feedback and consensus on the study topics includes all authors listed alphabetically from Ben Ashby to Daniel A. M. Villela, together with Mirjam E. Kretzschmar, Michael Mäs, Kai Nagel, Sydney Paltra, and Viola Priesemann of the infoXpand consortium, and Stefan Flasche, who is acknowledged below. Six infoXpand consortium members responded in Round 1, five of them panellists; the rest of the panel took part from Round 2 onwards, and 41 panellists completed the definitive Round 4 rating, as reported throughout.

Abstract

Infectious disease modelling shapes pandemic policy. How that modelling should and should not be done remains untested for agreement. We combined a systematic review with a modified Delphi study

among 41 experts in modelling, epidemiology, public health, and social sciences. The experts reworked statements drawn from the review, then rated them. All 40 statements reached consensus, and endorsement was strongest for 38 of them, at the highest possible median. Consensus was strongest on timely data availability, global modelling cooperation, and clear communication of model results and their scope of validity. However, by our assignment, modelling groups can act on only 18 recommendations independently. The rest need agencies, funders, or policy advisors to act between pandemics, demanding sustained funding, institutional mandates, and coordination beyond modelling. The panel worked primarily in high-income countries, so we offer these recommendations and ten priorities as an opening to a global discussion on improving pandemic preparedness.

Infectious disease modelling is now routine in outbreak response and has informed policy for decades. The 2001 foot-and-mouth disease epidemic in the United Kingdom is often cited as the first large-scale use of real-time modelling for that purpose.[1] Modelling has since informed responses to SARS,[2] pandemic influenza,[3] Ebola,[4] and COVID-19.[5] Across these responses, modelling analyses have helped policymakers estimate transmission, project healthcare demand, compare interventions, and allocate resources under uncertainty.[6]

COVID-19 brought this role under unprecedented scrutiny. Modelling informed decisions on lockdowns, physical distancing, vaccination, and resource allocation across many countries.[5,7–10] The pandemic also exposed weaknesses: limited transparency in assumptions,[11] poor communication of uncertainty,[12] fragmented data systems that hindered real-time calibration,[13] and strained relationships between modellers and policymakers.[14,15] Similar problems had surfaced in earlier outbreaks, including the 2009 H1N1 pandemic,[3] but COVID-19 made them far more visible and politically consequential.

A growing literature has since examined how modelling can better support public health decision-making.[16,17] It returns repeatedly to a few concerns: whether models and their assumptions are open enough to be scrutinised,[11,18] how uncertainty should be quantified and conveyed,[12,19] whether the necessary data exist and are governed well,[20,21] and how modelling connects to the decisions it is meant to inform.[6,22] Much of that learning appeared in print while the pandemic ran its course, and no one has consolidated it in full, so the advice remains scattered and untested for agreement.

We designed this study to consolidate that advice and determine what the modelling community agrees on. A systematic review supplied the evidence. Across the 159 publications that passed its initial screen (135 after we later tightened the criteria), modellers and the people who work with them set out the recurring challenges of modelling for outbreak response and what they proposed to address them. We turned that material into statements that a panel could rate, and a four-round Delphi study tested their agreement. We address the result to modelling groups, public health agencies (read broadly to include bodies that fund modelling or run national data infrastructure), and policy advisors who commission and interpret modelling work.

The panel is not the consortium that initiated the study. We opened recruitment within the infoXpand consortium and, from Round 2 onwards, invited additional experts through professional networks. The panel therefore grew as the study progressed. In all, 41 experts in modelling, epidemiology, public health, and social sciences completed the final round, and they are the panel we report throughout this work. They work primarily in high-income countries and in academia, which shapes the perspective these recommendations reflect.

The panel rated 40 statements across four sections: (A) model design and complexity; (B) data availability and quality; (C) uncertainty, validation, and communication; and (D) collaboration, interdisciplinarity, and ethics. It endorsed all 40, with 38 at the top of the rating scale, and that agreement fell more on the conditions surrounding a model than on how one should be built. We distinguish the items the panel rated, termed statements, from those that reached consensus, which we call recommendations. Since all 40 met the criterion, we carry every statement forward as a recommendation. Because that agreement concentrates on the conditions around a model, acting on most of these recommendations requires an agency, a funder, an infrastructure provider, or a policy advisor to work alongside the modelling group. Preparing for the next pandemic takes more than better models. It takes institutions funded between pandemics and modellers who say what their outputs mean and make the uncertainty in them accessible.

Results

Evidence base from the systematic review

The study ran in two stages. First, we used a systematic review to establish what the field has already said about modelling for outbreak response. Second, we conducted a four-round Delphi study in which an expert panel reworked and rated statements drawn from the review. Both stages are depicted in Fig. 1, and all 40 statements are given in Table 1, in the wording the panel rated. The table also shows the agreement each statement reached, along with our own judgement of which party must act on it. Each identifier gives the section (A to D, as headed in Table 1) and the statement's position within it.

After screening and reassessment, the review included 135 publications (see Fig. 2). These publications span the years 1989 to 2025; a majority of 90 (67%) were published from 2020 onwards, and reviews and systematic reviews account for 58% of the corpus. Most of the concerns already predate the COVID-19 pandemic. Among the 45 publications before 2020, *Model development and methodological improvement* appears in 64% and *Data governance, sharing, and infrastructure* in 56%. This is similar to their share among the 90 published since 2020 (58% and 47%). Only the *Reporting standards and documentation* theme is genuinely post-pandemic, absent before 2020 and present in 14% of cases thereafter. Supplementary Notes 1 and 2, with Supplementary Tables 1 and 2, break down publication type, period, and relevance, and list every included publication.

Eight themes in the reviewed literature

The review derived 568 recommendation items and 583 challenge items across the 159 publications of the initial screen. The synthesis of these items identified eight recurring clusters (see Supplementary Note 3 and Supplementary Table 3). *Model development and methodological improvement* dominates, appearing in 60% of the 135 publications, reflecting a literature written by modellers for modellers. Setting it aside, *Data governance, sharing, and infrastructure* is the most frequently raised topic at 50%, and *Reporting standards and documentation* the least, at 10%. We drafted the statements before this synthesis, so Table 2 aligns themes with statements retrospectively.

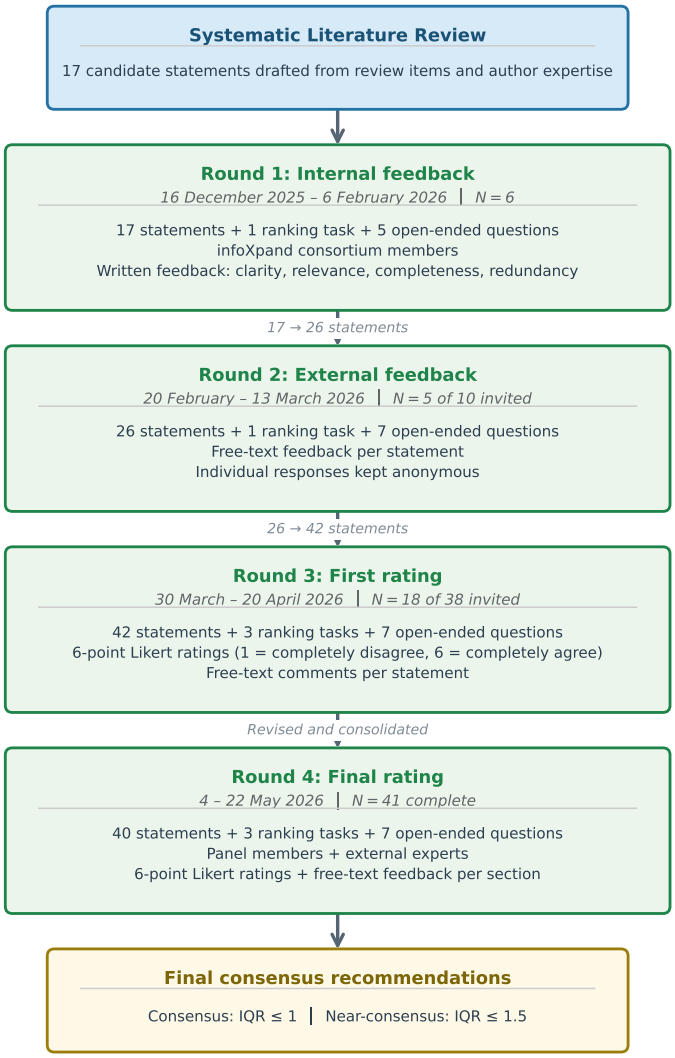


Figure 1: **Procedure of the study.** The study ran from the systematic review through the four Delphi rounds to the final consensus recommendations. Numbers give the respondents (N) and statements at each stage. The dates are the stated field periods for each round. Two Round 4 responses arrived after the survey's stated closing date, on 25 May and 5 June 2026, and both are retained in the analysis.

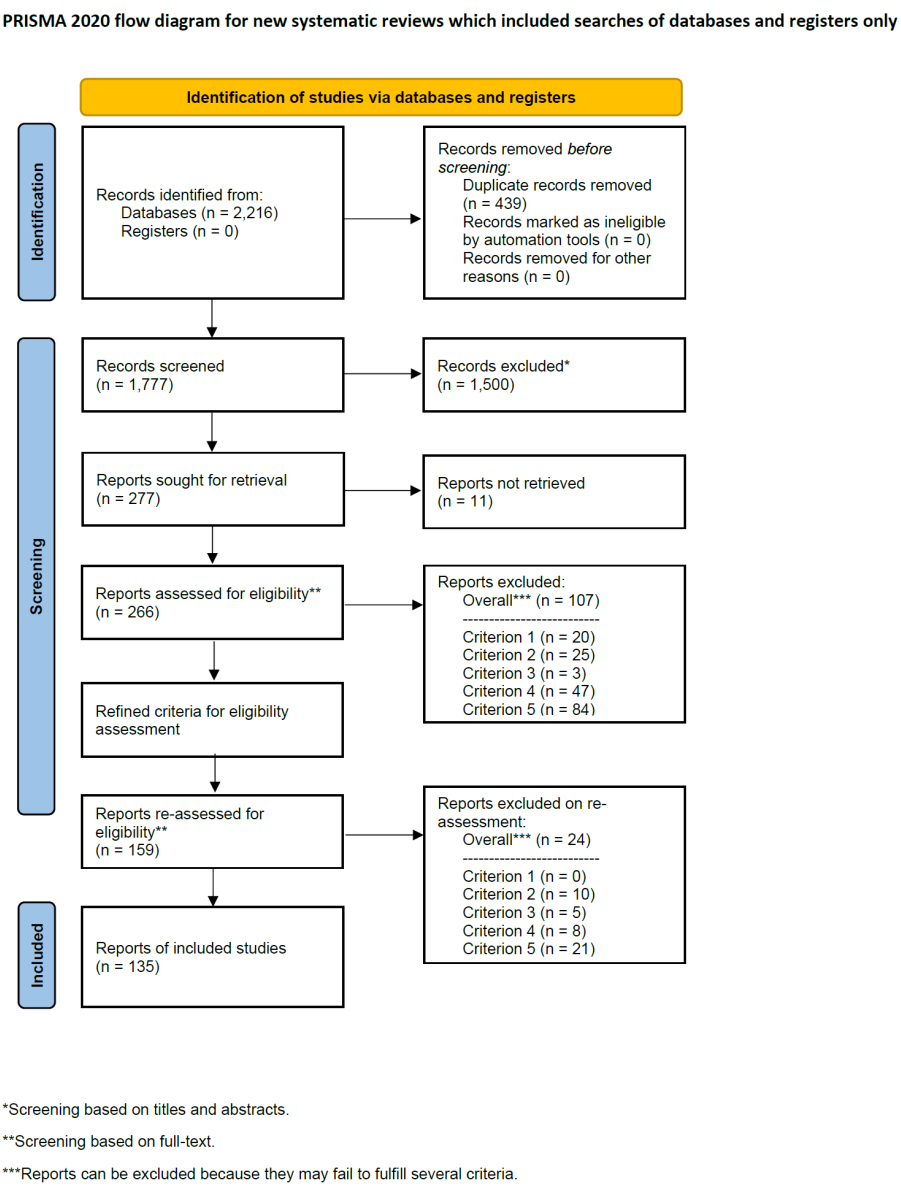


Figure 2: **Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 flow diagram for the systematic review.**[23] Boxes give the records identified, screened, retrieved, and included at each stage, with reasons for exclusion.

Consensus across all 40 statements

Agreement at the final rating was strong throughout. The six-point scale carried no neutral midpoint, and "Not qualified to respond" counted as an abstention excluded from n. Of the 40 statements, 38 had a median of 6, and all 40 met the consensus criterion of an interquartile range (IQR) no greater than 1. Not all 40 are actions: A02 and A08, and the opening sentences of D07 and D08, state propositions rather than steps, so agreement there records a shared reading of the problem rather than an endorsed action. At that ceiling, the median and IQR no longer separate the statements, so Table 1 also reports the top-two share, the percentage choosing 5 or 6, which ranges from 77% (C09) to 100% (D02), with a mean of 92% across all statements. Supplementary Note 4 and Supplementary Table 4 give the Round 3 comparison, and Extended Data Fig. 1 the round-on-round changes. None of our consensus classifications changes when restricting the sample to the 41 complete responses, compared with using partial responses. This only shifts one median slightly without changing the classification (A10, 5.5 to 5; Supplementary Note 5 and Supplementary Table 5). The number of experts invited to each round and the number who responded is shown in Fig. 3.

An agreement this strong could mean the statements were too general to argue with. Three features of the process say otherwise. First, consensus was not immediate: at Round 3, only 21 of the 42 statements then in the set met the criterion, and dispersion was greatest on multiple sampling approaches (B07, IQR 4). Panellists dismissed vague suggestions, such as "sensitivity analyses should be conducted carefully", as unhelpful. The revision therefore made the statements more operational, introducing predefined criteria, governance requirements, and documentation standards (Supplementary Note 6 and Supplementary Table 6). Second, the statement set grew from 17 to 40 as panellists added contested topics rather than dropping them, including child well-being and sentinel sampling. Third, free-text comments continued to record reservations at the final round, even where numerical agreement was already high. Supplementary Note 7 provides additional qualitative feedback.

Panellists also ranked seven preparedness measures within each of three pandemic phases (N = 45–46 per phase), stating importance rather than agreement. *Real-time data sharing* and *rapid-response modelling teams* led the pre-pandemic and early-emergence phases, while *empirical model validation* rose to first after the event (mean rank 3.09) and *rapid-response teams* fell to last (5.52), so validation matters once an outbreak has supplied data to

validate against. The panel held that post-event ordering least consistently, so we read it as indicative only (Supplementary Note 8 and Supplementary Table 7).

Strongest agreement on reporting and infrastructure, weaker on model construction

All 40 statements reached consensus, but the dispersion varied. Of the 40 statements, 13 had an IQR of 0, so at least three-quarters of raters chose the highest option. These name practices rather than goals. Six statements describe how modelling work should be conducted and reported, and a modelling group can implement all six without outside support:

- explicit calibration reporting (A13),
- uncertainty specified and quantified by type (C02),
- model quality judged against the question asked (A01),
- insight from integrating structurally different models (A03),
- speed paired with explicit caveats in rapid response (A05),
- and bias identified and mitigated (D06).

The other seven statements address the infrastructure and cooperation around a model. A modelling group cannot deliver these on its own. They require public health agencies, infrastructure providers, and policy advisors to act alongside the modelling group:

- timely data with documented limitations (B01),
- public data through free application programming interfaces (APIs) in machine-readable form (C08),
- stable, versioned, documented interfaces to reach it (C07),
- cross-border data-sharing agreements (D03),
- global cooperation and capacity building (D02),
- long-term interdisciplinary collaboration (D01),

- and ethical frameworks for data inequity (B08).

None of these 13 high-consensus statements prescribes a specific model class or implementation. Endorsement was weaker on statements about model design and construction. Both statements that fell below a median of 6 sit in Section A and address model complexity. They are concerned with the *strengths of agent-based models for scenario modelling* (A11, median 5) and *integrating socio-economic and behavioural factors* (A10, median 5.5). Both still met the criterion at an IQR of 1, placing their first quartiles at 4 or above, so at least three-quarters of raters agreed on a scale with no neutral option. Qualitative feedback matched this ambivalence, endorsing the statements while raising concerns about feasibility under crisis conditions and about the role of ethics in modelling.

Ten priorities for the inter-pandemic period

We distilled ten actionable priorities for the period between pandemics from the consensus recommendations and the review (Table 3).

All ten priorities ask for preparation rather than improvisation: fix the protocol, fix the framework, or fix the collaboration before the emergency arrives. A clear example is *interoperable real-time surveillance*: systems that exchange data through documented, versioned interfaces, using shared definitions and formats. This allows data collected by one agency to be combined with data from other agencies without extensive manual effort. *Cross-border agreements* settled in advance make that possible.

Our recommendations are divided according to who must act on them (Table 1): a modelling group can act on 18 of the 40 on its own, while 21 need an agency, a funder, or an infrastructure provider, and one needs a policy advisor rather than an agency (C09). Policy advisors have duties across three statements (B08, C09, D03). The same division runs through the ten priorities: two lie wholly within a modelling group's control, eight also require a public health agency, and four require a policy advisor.

Discussion

Overall, the Delphi panel reached high agreement (consensus) on all 40 statements. We divided the 13 statements with the strongest agreement by who

can act on them. A modelling group on its own cannot control public data infrastructure and collaboration, which seven statements address. It can control how modelling work is conducted and reported, which the remaining six address. Therefore, the panel agreed the most on aspects that are beyond model construction and often outside a modelling group's control. The panellists did not merely agree. They also rated the statements highly. Of the 40 statements, 38 reached the highest possible median, yet endorsement varied within the consensus.

One challenge during epidemics, the panel agreed, is data availability: where a model's data come from and how and when they reach the modelling group. To accelerate data availability, data-sharing protocols should be pre-negotiated and readily available (B03). Pre-negotiation removes the need to identify a counterparty and negotiate access under emergency time pressure. Data collection should be conducted using multiple sampling approaches (B07). Different sampling strategies (community-, hospital-, and wastewater-based) are needed to provide a holistic picture that is both timely and accurate. However, data collection is never free from bias, which should be transparently communicated and accounted for in analysis. Beyond what the panel rated, we would add that frontline services implementing policy decisions, such as family physicians and local public health units, should be systematically included in data collection and explicitly modelled. This group can provide uniquely valuable insights on working knowledge, contact patterns, and the practical effects of policy implementation (Supplementary Note 9).

Our panel agreed that the role of modelling in policy decisions should be strengthened (A12, D01), as the literature also argues.[8] This should close the gap between model development and policy implementation[20] to reduce reliance on the improvised advisory structures seen during COVID-19.[14,15,24] Yet, both the literature and the panel drew the same line: models advise; they do not prescribe. Few modellers or advisors fully master both modelling and policy, so dedicated bridging roles may work better than expecting either side to close the gap (D04).[17]

The panel agreed not only on transparency as an overarching goal in modelling infectious diseases but also on concrete reproducibility standards, as set out in the following three statements:

- explicit calibration reporting (A13),
- sensitivity analyses accompanying every policy-informing output (C05),

- and models, data, and code shared under the findable, accessible, interoperable, and reusable (FAIR) principles (B02).

Access to the data is most often missing but essential for reproducibility, since code reproduces a result only if the data it used remain retrievable, published in machine-readable form (C08) behind stable, versioned, documented interfaces (C07). In summary, our recommendations go beyond a mere call for openness to the conditions that make transparency achievable under time pressure.[11] They demand specific, checkable actions for transparency.

The panel agreed on the importance of communicating uncertainty (C01, C02, C09, C10), which the literature distinguishes from quantifying it.[12,19] While quantifying uncertainty is traditionally done by modellers, our panel placed the burden of communicating it comprehensibly on modellers rather than on their audiences (C09). C09 drew the lowest top-two share in the set at 77%, so this is where the panel was least united. Moreover, it broadened those audiences from policymakers alone to include the public and the media (C10).

The panel's central message was that models are tools for structured reasoning under uncertainty, not predictive oracles. Forecasts and scenarios sit on a continuum rather than either side of a strict dichotomy, as the panel endorsed (C03). Trouble begins when an output meant to explore what might happen is read as a forecast of what will happen. A scenario that does not materialise erodes confidence in modelling, even when it has served its exploratory purpose. Policy responses alter the trajectory that a scenario might project. This means that pre-intervention scenarios look wrong in hindsight precisely when they inform action, which makes them self-negating projections. Spontaneous pathogen evolution and behavioural change are also inherently hard to predict. An evaluation of the US COVID-19 Scenario Modeling Hub found that its projections tracked observed trajectories for an average of 22 weeks, until SARS-CoV-2 variants that no scenario had anticipated broke the assumptions underpinning them.[25] A scenario the world did not follow is, therefore, not evidence that the model failed. It shows instead that a model's validity is constrained by its assumptions and that models should not be applied beyond their scope.

The two statements with slightly lower median ratings, both concerning model complexity, still met the consensus criterion. Here, panel comments raised the question of whether the added complexity is justified when the additional parameters cannot be estimated with sufficient precision. Where

the obstacle is computational cost rather than estimation, machine-learning surrogates trained on mechanistic models offer one way to keep detail affordable, at the price of transparency.[26] Behavioural parameters, inherently heterogeneous and hard to measure precisely, could be elicited through novel means such as interactive games.[27] Forecast hubs make the ensemble case: their multi-model approaches were more consistently accurate than any single model.[28] For scenario projections, pooling models without collapsing their disagreement proved more dependable than any individual model, so long as the scenario's assumptions still held.[25] Preserving those differences is, therefore, essential to quantifying structural uncertainty. The question is not whether complex models are worth developing, but when they justify their cost.

Not every choice in building a model is technical. Which populations a model represents and which outcomes it collects are value choices. Deciding not to model a societal group renders that group invisible to the policy the model will inform. The panel asked that those decisions be examined as a model is built, with attention to resource allocation and vulnerable groups (D05).

Our recommendations complement the agenda announced for the Lancet Commission[8] and earlier initiatives on transparency,[11] uncertainty,[12] data,[20] and the science-policy interface.[29–31] They sit alongside instruments that govern narrower ground: the EPIFORGE reporting checklist, which sets out what epidemic forecasting studies should report,[18] and the framework the World Health Organization lays out, which guides countries in strengthening preparedness capacity.[32] By contrast, we address the institutional conditions that enable good reporting.

Every priority (Table 3) names the party that must act alongside the modelling group, and for all but two, that is somebody else. Transparency, documented assumptions, and reporting standards are matters of professional practice, whereas interoperable platforms, multi-model hubs, and cross-border agreements presuppose infrastructure and sustained funding that many settings lack (Supplementary Note 10). For this reason, global capacity building and equity matter. Pandemic threats fall most heavily on low- and middle-income countries (LMICs), while modelling capacity and research output are concentrated far from those settings.[7,33] Global modelling cooperation (D02) drew some of the strongest endorsement of any recommendation.

Much of this has been said before, and the systematic review documents

how long ago it was said. The topics of *Data governance, sharing, and infrastructure* and *Model development and methodological improvement* were raised at least as often before 2020 as after 2020, and the earliest publications from the review already argue for closer coupling between models and policy. Progress on both needs funding and a mandate to act between emergencies, when public attention has moved on.

Written largely by modellers, these recommendations describe an arrangement we consider ideal rather than one negotiated with the parties who would have to build it. The counterpart, what agencies need from modelling groups, is the study to run next. Side by side, the two would expose where expectations conflict (Supplementary Note 10).

The panellists also raised pressing questions that our study has not settled. One of the questions was how to integrate behaviour and social dynamics into models and to broaden outcomes beyond acute morbidity. One Health, which couples human, animal, and environmental health, was another, as was the integration of biological knowledge. Neither drew much attention from the panel or the literature, though both are central to preparedness for emerging threats. While transparency was endorsed by the panel, the practice of pre-registering modelling studies was discussed ambivalently. Since pre-registering conditional assumptions risks framing outputs as predictions to be scored, it creates negative incentives for scenario modelling. Regarding uncertainty, the panel named the audiences but did not settle how it should be tailored to each, a question that still needs further investigation in this domain. One overarching question lingered: whether advisory arrangements actually affected decisions during COVID-19.[8,14,15,30] Future efforts should address these topics and seek insights on endorsement and consensus where possible.

Some limitations remain. The first limitation is who produced these recommendations. Of the 41 final-round panellists, 36 work in Europe and 40 are academics. The panel is therefore drawn almost entirely from high-income academic settings and speaks predominantly to those who produce modelling evidence rather than those who apply it (Supplementary Note 11). Recommendations describing professional practice should hold wherever modelling informs policy, while those that presuppose infrastructure will transfer less readily. The same study run among modellers in LMICs would establish how far these recommendations travel.

Second, these recommendations also have gaps. Health economics is the clearest. No statement addresses cost-effectiveness or budget impact directly,

and economic evaluation survives only obliquely in A04 and A12, even though the review returned work on it and the panel includes health economists. Furthermore, implementation is treated as simpler than it is. The recommendations presuppose a functioning state and a health system with the authority to act. They say little about the political contestation in which implementation usually fails. Both gaps mark the limits of what these recommendations cover.

Lastly, the method carries its own limits. Because the panel broadened across rounds, the results are agreement within the final panel rather than convergence of a fixed one. Round 4's anonymity leaves its retention status unknown, and no check here can recover it. Two further limits have checks: a completers-only reanalysis tests whether the unsubmitted responses shift the results by changing who is in each denominator, and the top-two share separates statements once the median saturates at the top of the scale (Supplementary Note 5). The eight themes are one defensible organisation of the material, and the codebook behind them agrees only moderately with hand labels. Yet, neither the themes nor the codebook informs the consensus results, which predate both. Our relevance framework remains unvalidated (Supplementary Note 12). Consensus does not establish correctness; ratings can be pulled towards group medians. We did not ask the panel to weigh the recommendations against one another, so where they compete for funding or attention is beyond what these data settle. COVID-19 dominated the literature reviewed, and we included only English-language publications. In our judgement, none of these limitations undermines the consensus results or the recommendations drawn from them.

Two things follow from our recommendations.

First, what surrounds a model needs at least as much attention as the model itself. The panel's strongest endorsements fall on problems the field has named for two decades. What is new is that the panel has stated them in terms specific enough to act on. The remaining gap is one of implementation rather than knowledge. The recommendations needing someone beyond the modelling group ask for something concrete:

- machine-readable, public data behind documented, versioned interfaces (C07, C08),
- pre-negotiated data-sharing protocols and cross-border agreements (B03, D03),

- joint modelling hubs with capacity building in under-resourced settings (D02),
- standing interdisciplinary collaborations (D01),
- and funded roles that bridge modelling and policy (D04, D09).

What these items need is sustained funding and an institutional mandate, not further scientific agreement.

Second, where modellers bear responsibility, they should publish their outputs and explain what those outputs mean (A13, C01 to C03, C05, C09, C10):

- the assumptions they rest on,
- their scope of validity,
- whether they forecast what will happen or explore what could happen under stated assumptions,
- and the associated uncertainty, expressed as ranges rather than point values.

We would add that the ranges that modellers published during COVID-19 often reached the public as single numbers, so science communicators bear part of the responsibility, too.

The pandemic assembled a modelling community faster than it built the institutions that community needed. These recommendations are an attempt to leave behind something durable: a common language in which such proposals can be stated, compared, and revisited as evidence accumulates. We offer them as an opening position rather than a settled one, published to start a process that will need other voices to finish. Our recommendations should be addressed now rather than when the next pandemic arrives.

Methods

Study design

The infoXpand consortium, a multidisciplinary research network, ran the study. We conducted a systematic review and used it to inform a modified

Delphi process to develop expert consensus on recommendations for infectious disease modelling in pandemic preparedness. The Delphi comprised four sequential rounds (Fig. 1): Rounds 1 and 2 developed the statement set through qualitative feedback, and Rounds 3 and 4 rated it quantitatively. The panel broadened at each round, so Round 4 measures endorsement of the final set by the widest panel we could assemble rather than convergence within a closed group.

Systematic review

We report the search and selection using the PRISMA 2020 flow diagram.[23] We searched PubMed, Scopus, and Web of Science on 7 February 2025 using a structured query combining disease context, modelling approach, public health relevance, and critical reflection. Publications were eligible against five criteria: they (1) addressed human infectious disease outbreaks, (2) discussed a model-based approach, (3) were relevant to public health, (4) critically examined challenges, and (5) took a broader perspective on modelling practice. We further restricted eligibility to publications available in English and did not log language as an exclusion reason on its own, so Fig. 2 includes non-English records in its screening counts rather than showing them as a category. Supplementary Note 13 states the language restriction alongside the five content criteria and gives operational definitions of criteria (2) to (5) with worked examples, and Supplementary Note 14 provides the search strings.

The review follows systematic review methodology with three adaptations: one author screened titles and abstracts; we assessed relevance in place of risk of bias, since no validated tool fits a corpus spanning commentaries to research articles; and we pre-registered the Delphi analysis plan (https://osf.io/p7wm3) rather than a review protocol.

Screening and selection

The search returned 2,216 records. After deduplication, title and abstract screening, and full-text assessment of the 266 reports we retrieved, 159 publications passed the initial screen, and we developed the Delphi statements from that set.

We later tightened the eligibility criteria and reran the screen. This changed which publications the review reports on, not which evidence the

Delphi drew on: we had already written the statements from the original 159, and every rating predates the reassessment. Revisiting the screen, we found that criteria 4 and 5, which drove most of the decisions, rested on judgement rather than on a stated test. We defined explicit tests for both (Supplementary Note 13) and reassessed all 159, excluding 24, most often under criterion 5 (21 publications), so the review reports 135.

One author (H. Nunner) screened titles and abstracts single-handedly, a pragmatic choice given the volume of records. The title and abstract screening was, therefore, not independently duplicated.

The full-text assessment used a structured prompt-based protocol, with OpenAI GPT-4o mini for the initial screen and Anthropic Claude Sonnet 5 for the reassessment. Given the documented weaknesses of these models, we treated every LLM output as a first pass, and the research team owned all inclusion, exclusion, and classification decisions. Systematic verification against the source articles was limited to the reassessment, and the outputs of the initial screen and extraction were not checked against the sources individually (Supplementary Note 15). One author (L. Stellbrink) verified 84 of the 159 reassessed records (53%), including every proposed exclusion ($n = 24$) and a randomly selected subset ($n = 60$) of the included publications, and overturned none of them. Supplementary Note 15 gives the protocol, the prompt, and the full accounting.

Relevance framework

The review spans commentaries and research articles, so no validated quality appraisal tool applies to it. Our first attempt at one failed: a six-criterion rubric placed 96% of publications in a single band and so discriminated against nothing. The problem was that it scored eligibility and relevance together. Separating them fixed it, and the resulting bands spread across high, moderate, and low relevance (Supplementary Table 1). Eligibility is determined by the five criteria above, each of which is either met or not met. Relevance is scored on four dimensions, each rated from 0 to 2, and captures the breadth of discussion rather than methodological quality. Supplementary Note 12 and Supplementary Table 8 provide the dimensions, wording at each level, and band thresholds.

Data extraction

The same LLM-assisted protocol extracted key findings, challenges, and recommendations from the 159 publications of the initial screen. These items informed statement drafting and the thematic synthesis. Of these 1,151 items, 987 came from publications that also met the sharpened criteria, and the 24 excluded on reassessment fell outside the sharpened scope rather than short of quality.

Delphi panel and recruitment

We recruited the panel through the infoXpand consortium, selecting members for expertise evidenced by publications, advisory roles, or direct involvement in modelling for policy. We invited them by personalised email, first within the consortium and then, in Rounds 2 to 4, externally. We sought geographic and disciplinary diversity, but recruitment centred on Europe, with additional invitations in North America, Oceania, and Latin America, reflecting the consortium's institutional base. Extended Data Table 1 presents the composition of the final-round panel, which includes the 41 who submitted and excludes the additional raters behind each per-statement denominator, so up to 10 raters per statement have unknown region, sector, and experience. It records sector and career stage rather than discipline, so the disciplinary breadth claimed for the panel rests on the recruitment criteria rather than on a tabulated count. Supplementary Note 11 describes recruitment and composition, and Supplementary Note 16 describes the round-by-round procedures.

Typically, Delphi panels report the attrition rate to contextualise the consensus process. Our panel was broadened deliberately, so the rate of attrition is not easily measured. Instead, we report each round's response rate and, when applicable, the retained participants from the previous round. Fig. 3 separates invitations from responses and splits respondents into those retained and those responding for the first time. Rounds 1-3 used a personalised invitation process, so retention is determinable. Round 4 was anonymous. Not all authors rated the statements; only panel members did, so every Round 4 response comes from a panel member or an external expert. We distributed Round 4 as a shareable link and encouraged forwarding to suitable colleagues, so we cannot say how many respondents arrived through a forwarded link. Responses were anonymous; thus, the final ratings include

an unknown number of respondents whose identity we could not verify. Supplementary Notes 11 and 16 describe the procedure of sharing the link and how to interpret the self-reported characteristics of the final panel.

Statement generation

Statement drafting drew on the items extracted from the initial 159 publications, a team-authored document of candidate ideas from the review, our own modelling projects, and practitioner experience. A project-specific GPT-5 agent produced a first draft from those inputs, and the team then reviewed, consolidated, and rewrote it against the reviewed literature. Accordingly, the statements are evidence-informed rather than evidence-derived. The Round 1 set comprised 17 statements, already divided into the four sections (A) to (D) used throughout (Supplementary Note 17 and Supplementary Table 9). It also included a ranking task of seven pandemic preparedness measures and five open-ended questions. Feedback expanded it to 26 for Round 2 and 42 for Round 3. After Round 3 we merged two pairs of statements, giving the 40 rated in Round 4. Statement identifiers (e.g., A11) always refer to the Round 4 set. Supplementary Note 6 and Supplementary Table 6 map the revisions and the Round 3 numbering.

Thematic synthesis

The eight themes are a retrospective description of the literature, not a source of the statements. We synthesised them only after we finalised the statement set, using an LLM (Anthropic Claude Opus 4.6) on the review workbook alone, without panel input. Table 2 maps the eight themes onto the statements. Because the thematic synthesis did not record which themes each publication addressed, we re-derived coverage for the 135 publications using an explicit list of search terms (Supplementary Table 10). We validated it on a second random sample of 30 publications hand-labelled by one author (L. Stellbrink), reaching moderate agreement (Cohen's $\kappa = 0.55$), precision 0.82, and recall 0.62. Five of the 30 had appeared in the tuning sample, so they are not strictly held out, and excluding them yields $\kappa = 0.53$ (Supplementary Note 18). Because one author produced the reference labels, κ measures the codebook against a single-coder standard rather than between two independent coders. Supplementary Note 18 gives the method.

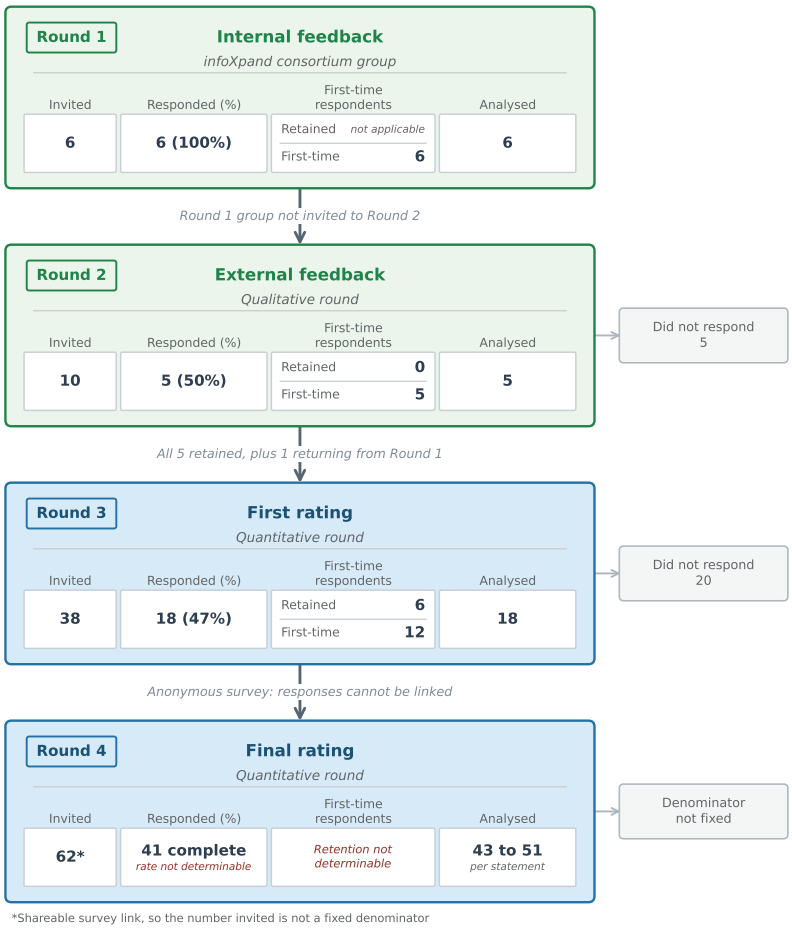


Figure 3: **Panel flow across the four Delphi rounds.** Every round reports the same four quantities, and a cell states where one does not apply or cannot be determined. *Invited* counts the experts asked to take part, and *Responded (%)* how many answered and their share of those invited. *First-time respondents* splits them into those retained from an earlier round and those answering for the first time, which sum to *Responded (%)*. *Analysed* counts the responses that entered the analysis. Round 2 went only to external members, so no Round 1 panellist carried over; from Round 3 we re-invited everyone previously invited unless they had declined. Panellists forwarded Round 4 beyond its 62 direct invitees, so its denominator is not fixed, its response rate is only approximate, and its retention is indeterminable. Its analysed counts are raters per statement, since partial responses were kept.

Rating and response handling

In the two quantitative rounds, Round 3 ($N = 18$) and Round 4 ($N = 41$ submitted responses), panellists rated each statement on a six-point Likert scale (1 = completely disagree, 6 = completely agree), completed three phase-specific ranking tasks, and answered seven open-ended questions. The even-numbered scale avoided a neutral midpoint. Round 4 also offered "Not qualified to respond", which we treated as an abstention throughout: it enters no median, interquartile range, or share, and does not count towards n. Thirty-six abstentions fell across 19 of the 40 statements, listed per statement in Supplementary Note 19. Between 47 and 51 panellists answered each statement by rating or abstaining, so abstention, rather than partial nonresponse, accounts for much of the spread in n. We followed the pre-registered analysis plan, available at https://osf.io/p7wm3, with one documented deviation regarding repeat submissions. The statements have between 43 and 51 responses, because we retained partial responses (Table 1). We report $N = 41$ complete responses, and call this our panel size. We define the panel by those complete responses because the final question is where a panellist confirmed the response is theirs. Supplementary Note 19 describes response handling and reconciles every Round 4 count reported in this article.

Consensus criterion

We defined consensus a priori by the interquartile range: consensus at IQR ≤ 1, near-consensus above 1 and up to 1.5, and non-consensus above 1.5, applied to Round 4 as the definitive assessment.[34–36] We registered the analysis plan, including this criterion, on 23 April 2026, after Round 3 closed and before the definitive Round 4 rating opened on 4 May 2026. The criterion was therefore fixed before the ratings it classifies, while the Round 3 figures reported alongside them are a retrospective application of it. Statements that failed in both rounds would have been counted as stable non-consensus.[37] Quartiles use linear interpolation, so medians and interquartile ranges can take non-integer values such as 5.5 or 0.75 (Supplementary Note 5). Because median and IQR saturate once agreement is high, we also report the top-two share, the percentage of raters choosing 5 or 6, which still separates statements that the median and the IQR no longer do. We ran two checks on the consensus results. Restricting every statistic to the 41 complete responses tests whether submission status drives the outcome, since it drops every un-

submitted response, including the six that answered all 40 items, and the top-two share shows the variation the ceiling hides; Supplementary Table 5 reports both per statement. Supplementary Note 19 describes the analysis pipeline.

Assigning who must act

The panel rated the statements but was not asked who should act on them. We made that assignment ourselves afterwards by reading each statement (Supplementary Note 10), so the "must act" column reflects our judgement rather than the panel's. We applied one rule: a party bears a label when the statement's text places a duty on it, or when a modelling group cannot enact the statement without it, and not when the party appears only as the recipient of communication. The rule identifies the party that must act, which is not always every party that takes part. A reader who counts participants rather than initiators will divide the set differently, and Table 1 gives the statement wording needed to do so. The coding was a team judgement made in one pass and was not independently duplicated, and we report no agreement statistic for it. We grouped the ten priorities of Table 3 by what they ask for and which parties must act on them, again, our judgement rather than the panel's.

Use of large language models

We used LLMs as supervised assistants for well-specified tasks. They produced:

- the preliminary full-text screening, relevance scoring, and structured extraction for the review (OpenAI GPT-4o mini),
- the eligibility reassessment (Anthropic Claude Sonnet 5),
- a first draft of the statement set (an agent built on OpenAI's GPT-5),
- and the eight-theme synthesis (Anthropic Claude Opus 4.6).

We synthesised panellists' free-text comments with a locally hosted DeepSeek-V4 model, so no panellist-authored text left institutional control. The authors made all inclusion, exclusion, relevance, thematic, and statement-wording decisions and take full responsibility for the content.

In preparing the manuscript, the authors used Grammarly, Quillbot, and Anthropic Claude to check language, style, and structure, then reviewed and edited the results. We did not log exact model snapshots, access dates, or decoding parameters for each run, so we list the model families above rather than version strings. Supplementary Note 15 identifies the model used at each step, assesses this reporting against TRIPOD-LLM, and provides the protocols, prompts, and the limits of the verification we were able to perform.[38]

Ethics

This study did not require formal ethics committee approval at the University of Luebeck. The university's ethics committee does not review Delphi studies of this kind, which collect no special-category data and pose no risk to the panellists. We therefore could not obtain a formal opinion, and none was required. Professional experts participated voluntarily, were fully informed of the study's purpose, affirmed a consent statement at the start of each round, and could withdraw at any time without consequence. We treated all individual responses as confidential and shared only aggregated results during the iterative rounds and in this publication.

Reporting summary

Supplementary Note 20 and Supplementary Table 11 report the systematic review against the 27 items of the PRISMA 2020 statement, item by item, including the items that a review of this kind cannot report and the three that we did not do.[23] We designed the Delphi against the CREDES guidance,[37] but no completed item-by-item checklist accompanies this article. Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The pre-registered analysis plan is available at https://osf.io/p7wm3. Supplementary Note 21 sets out the dissemination plan and how the set is to be kept up to date. The quantitative data underlying this study are deposited at Zenodo under a CC BY-SA 4.0 licence

(https://doi.org/10.5281/zenodo.22232256): the anonymised Round 4 rating data and the Round 3 coded output, the full-text screening export underlying the review, and the hand labels used to validate the term-based codebook.[39] The CC BY-SA 4.0 licence covers only the material we authored. The screening export quotes short passages from the reviewed publications, reproduced as evidence; copyright in those passages remains with their rights holders, and the licence we grant does not extend to them or permit their redistribution under share-alike terms. These files, when run through the analysis code below, reproduce the review tables, theme shares, and consensus statistics reported here.

Code availability

All custom code is available under an MIT licence. The Delphi analysis pipeline, the term-based theme codebook, and the scripts that build every computed table and figure are archived at Zenodo (https://doi.org/10.5281/zenodo.22679573) and developed at https://github.com/lstell21/pg-delphi-analysis.[40] The LLM-assisted full-text screening instrument, which holds the five inclusion criteria and the decision rule applied to them, is archived at Zenodo (https://doi.org/10.5281/zenodo.22647106) and developed at https://github.com/hnunner/pg-literature-screening.[41] Five display items are hand-made rather than computed, so the archive does not derive them: the PRISMA diagram (Fig. 2), the theme-to-statement map (Table 2), the ten inter-pandemic priorities (Table 3), the Round 1 statement set transcribed in Supplementary Table 9, and the PRISMA checklist in Supplementary Table 11. An archived script writes out Table 3, but the grouping and the wording in it are ours. The screening run reported here is not reproducible because it called a large language model and required full texts that we cannot redistribute. The deposited export is therefore the run of record, and every step downstream of it is deterministic.

Acknowledgements

The authors thank Daniel Ernst for research assistance with the initial relevance appraisal of the reviewed publications, conducted within a framework

that was later retired in favour of the approach reported here. The authors also thank Stefan Flasche for expert input as a member of the Delphi panel. With a heavy heart, we record that Anthony Staines, who played an important part in this study, died before the final editing was completed. We remember with gratitude his friendship and contributions.

Author contributions

L.S.: Conceptualisation, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Visualisation, Writing – original draft, Writing – review and editing. **H.N.:** Conceptualisation, Data curation, Software, Methodology, Visualisation, Investigation, Writing – review and editing. **A.S.:** Data curation, Investigation, Writing – review and editing. **M.M.:** Software, Writing – review and editing. **A.C.V.:** Conceptualisation, Methodology, Supervision, Writing – review and editing. **M.E.K., M.Mä., K.N., V.P.:** Conceptualisation, Funding acquisition, Investigation, Writing – review and editing. **S.P., B.A., P.B., T.B., S.C., T.C., U.D., E.F., N.M.F., F.G., E.G., R.G., E.Gr., C.H., N.H., T.H., M.J., A.K., P.K., T.K., A.Ku., M.Kü., J.L., N.L., C.M., J.M.M., M.McK., R.M., C.P., J.P.-G., D.P., L.P., M.P., E.P., M.Pi., M.J.P., N.P., B.P., J.R., F.S., H.S., A.St., B.S., E.S., R.N.T., M.T., D.A.M.V.:** Investigation, Writing – review and editing.

Competing interests

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper. The following authors declare the advisory and committee roles listed below as interests: T. House and E. Fearon are members of the Scientific Pandemic Infections Group on Modelling (SPI-M), a committee that advises the UK Department of Health and Social Care. J. M. McCaw is an invited expert member of the Communicable Diseases Network Australia, the Australian government's committee tasked with providing national public health coordination and leadership. M. McKee and C. Pagel are members of Independent SAGE, a group of volunteers that conducted public engagement during the pandemic. M. Pickersgill is a member of the Scottish Science Advisory Council. This

article was written in an independent capacity.

Funding

The following authors disclose support for the research of this work from the Federal Ministry of Research, Technology, and Space (BMFTR), Germany, through the infoXpand consortium: V.P. under the Göttingen sub-project [031L0300A]; M.Mä. under Karlsruhe [031L0300B]; L.S., H.N., A.S., M.M. and A.C.V. under Lübeck [031L0300C]; and S.P. and K.N. under Berlin [031L0300D]. M.E.K. discloses support from The Netherlands Organisation for Health Research and Development (ZonMw) via the project "Control of Infectious Diseases: modelling decisions and behaviour in Social Networks (ConnectioN)" [10710062310022]. N.M.F. discloses support from the UK Medical Research Council (MRC) Centre for Global Infectious Disease Analysis [MR/X020258/1], the National Institute for Health and Care Research (NIHR) Health Protection Research Unit in Health Analytics and Modelling, and a philanthropic donation from Community Jameel supporting the work of the Jameel Institute. R.G. discloses support from The Institute for Global Pandemic Planning at the University of Warwick, UK, as part of a philanthropically supported doctoral programme. P.B. and N.H. disclose support from the Belgian Pandemic Intelligence Network (BE-PIN), a project sponsored by the Belgian Science Policy Office (BELSPO), and from the Efficient and rapidly SCAlable EU-wide evidence-driven Pandemic response plans through dynamic Epidemic data assimilation (ESCAPE) project [101095619], funded by the European Union Horizon Europe research and innovation programme. T.H. and E.F. disclose support from the COMMET project [UK Research and Innovation (UKRI) grant MR/Z505328/1]. T.H. and L.P. disclose support from the Wellcome Trust [227438/Z/23/Z] and the MRC [UKRI483]. T.H., L.P., F.S., H.S. and R.N.T. disclose support from the JUNIPER partnership, supported by the MRC [MR/X018598/1]. A.K. discloses support from the BMFTR, formerly the Federal Ministry of Education and Research (BMBF), via OptimAgent [031L0299J], RESPINOW [031L0298F], VBD-MODE [031L0828E], TwinChain [031L0325A], PREPARED [100665502] and NUM-SAR [01KX2521], from the German Research Foundation (DFG) via EpiAdaptDiag [458526380] and ARCANE [530002115], from Interreg Deutschland-Nederland via CARE-FLOW [12156] and from the Volkswagen Foundation via the INFRALINK project.

M.Kü. discloses support from the BMFTR via TwinChain [031L0325A] and Episerve [031L0324E]. J.M.M. discloses support from an Australian Research Council Laureate Fellowship [FL240100126]. S.P. and K.N. disclose further support from MODUS-COVID [031L0302A], Episerve [031L0324B] and TU Berlin. J.P.-G. discloses support from the UK Health Security Agency, the UK Department of Health and Social Care, and the Oxford EPSRC Centre for Doctoral Training in Healthcare Data Science [EP/Y035321/1]. D.P. discloses support from the ESCAPE project [101095619], funded by the European Union Horizon Europe research and innovation programme. M.P. discloses support from the Slovenian Research and Innovation Agency [P1-0403]. M.J.P. discloses support from the Marsden Fund [24-UOC-020]. D.A.M.V. discloses support from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) [Finance Code 001], the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [307057/2026-7], and the Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) [E-26/204.108/2024]. Views expressed are those of the authors and do not necessarily reflect those of the European Union, the European Health and Digital Executive Agency, the UK Health Security Agency, the UK Department of Health and Social Care, or the Oxford EPSRC Centre for Doctoral Training in Healthcare Data Science. Neither the European Union nor the granting authority can be held responsible for them. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Keeling, M. J. *et al.* Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. *Science* **294**, 813–817 (2001).

2. Riley, S. *et al.* Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. *Science* **300**, 1961–1966 (2003).

3. Moghadas, S., Pizzi, N., Wu, J., Tamblyn, S. & Fisman, D. Canada in the face of the 2009 H1N1 pandemic. *Influenza and Other Respiratory Viruses* **5**, 83–88 (2011).

4. Thompson, R. *et al.* Using real-time modelling to inform the 2017 Ebola outbreak response in DR Congo. *Nature Communications* **15**, 5667 (2024).

5. Aguas, R. *et al.* Modelling the COVID-19 pandemic in context: an international participatory approach. *BMJ Global Health* **5**, e003126 (2020).

6. Jit, M. & Cook, A. R. Informing Public Health Policies with Models for Disease Burden, Impact Evaluation, and Economic Evaluation. *Annual Review of Public Health* **45**, 133–150 (2024).

7. Adib, K. *et al.* A participatory modelling approach for investigating the spread of COVID-19 in countries of the Eastern Mediterranean Region to support public health decision-making. *BMJ Global Health* **6**, e005207 (2021).

8. Lancet Commission on Strengthening the Use of Epidemiological Modelling of Emerging and Pandemic Infectious Diseases. How modelling can better support public health policy making: the Lancet Commission on Strengthening the Use of Epidemiological Modelling of Emerging and Pandemic Infectious Diseases. *The Lancet* **403**, 789–791 (2024).

9. Conway, E. *et al.* COVID-19 vaccine coverage targets to inform reopening plans in a low incidence setting. *Proceedings of the Royal Society B: Biological Sciences* **290**, 20231437 (2023).

10. Panovska-Griffiths, J. *et al.* Determining the optimal strategy for reopening schools, the impact of test and trace interventions, and the risk of occurrence of a second COVID-19 epidemic wave in the UK: a modelling study. *The Lancet Child & Adolescent Health* **4**, 817–827 (2020).

11. Round, J., Kirwin, E., van Katwyk, S. & McCabe, C. Improving Transparency of Decision Models Through the Application of Decision Analytic Models with Omitted Objects Displayed (DAMWOOD). *Pharmacoeconomics* **42**, 1197–1208 (2024).

12. Swallow, B. *et al.* Challenges in estimation, uncertainty quantification and elicitation for pandemic modelling. *Epidemics* **38**, 100547 (2022).

13. Akman, O. *et al.* The Hard Lessons and Shifting Modeling Trends of COVID-19 Dynamics: Multiresolution Modeling Approach. *Bulletin of Mathematical Biology* **84**, 3 (2022).

14. Hadley, L. *et al.* Challenges on the interaction of models and policy for pandemic control. *Epidemics* **37**, 100499 (2021).

15. Kretzschmar, M. E. *et al.* Challenges for modelling interventions for future pandemics. *Epidemics* **38**, 100546 (2022).

16. Adams, S., Lancaster, K. & Rhodes, T. Undoing elimination: Modelling Australia's way out of the COVID-19 pandemic. *Global Public Health* **18**, 2195899 (2023).

17. Hadley, L. *et al.* How understanding the diversity of perspectives and systems in governments can increase the impact of scientific research. *Communications Health* **1**, 13 (2026).

18. Pollett, S. *et al.* Recommended reporting items for epidemic forecasting and prediction research: The EPIFORGE 2020 guidelines. *PLOS Medicine* **18**, e1003793 (2021).

19. Fisman, D. N. When the role of uncertainty is... uncertain. *Proceedings of the National Academy of Sciences* **120**, e2305856120 (2023).

20. Cauchemez, S., Hoze, N., Cousien, A., Nikolay, B. & Ten Bosch, Q. How Modelling Can Enhance the Analysis of Imperfect Epidemic Data. *Trends in Parasitology* **35**, 369–379 (2019).

21. Avramov, M. *et al.* A conceptual health state diagram for modelling the transmission of a (re)emerging infectious respiratory disease in a human population. *BMC Infectious Diseases* **24**, 1198 (2024).

22. Ali, S. *et al.* Incorporating Social Determinants of Health in Infectious Disease Models: A Systematic Review of Guidelines. *Medical Decision Making* **44**, 742–755 (2024).

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

24. McCaw, J. M. & Plank, M. J. The Role of the Mathematical Sciences in Supporting the COVID-19 Response in Australia and New Zealand. *The ANZIAM Journal* **64**, 315–337 (2022).

25. Howerton, E. *et al.* Evaluation of the US COVID-19 Scenario Modeling Hub for informing pandemic response under uncertainty. *Nature Communications* **14**, 7260 (2023).

26. Schmidt, A., Zunker, H., Heinlein, A. & Kühn, M. J. Graph neural network surrogates to leverage mechanistic expert knowledge towards reliable and immediate pandemic response. *Scientific Reports* **16**, 6361 (2026).

27. Colubri, A. *et al.* Understanding human behaviour for pandemic preparedness with epigames. *Nature Health* **1**, 686–688 (2026).

28. Cramer, E. Y. *et al.* Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proceedings of the National Academy of Sciences* **119**, e2113561119 (2022).

29. Lee, J. *et al.* Institutional and behaviour-change interventions to support COVID-19 public health measures: a review by the Lancet Commission Task Force on public health measures to suppress the pandemic. *International Health* **13**, 399–409 (2021).

30. Jit, M. *et al.* Reflections On Epidemiological Modeling To Inform Policy During The COVID-19 Pandemic In Western Europe, 2020–23: Commentary reflects on epidemiological modeling during the COVID-19 pandemic in Western Europe, 2020–23. *Health Affairs* **42**, 1630–1636 (2023).

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

32. WHO. Framework for health emergency preparedness and response capabilities for national public health agencies (2026). `https://pandemichub.who.int/publications/framework-for-health-emergency-preparedness-and-response-capabilities-for-national-public-health-agencies`.

33. Adetokunboh, O., Mthombothi, Z., Dominic, E., Djomba-Njankou, S. & Pulliam, J. African based researchers' output on models for the transmission dynamics of infectious diseases and public health interventions: A scoping review. *PLOS ONE* **16**, e0250086 (2021).

34. 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).

35. 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).

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

37. 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).

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

39. Stellbrink, L., Nunner, H. & Schönberger, A. Data for "Preparing for the Next Pandemic Takes More Than Better Models: Expert Consensus Recommendations for Infectious Disease Modelling" (2026). `https://zenodo.org/doi/10.5281/zenodo.22232256`.

40. Stellbrink, L. Pandemic guidelines: Delphi analysis and review synthesis (2026). `https://zenodo.org/doi/10.5281/zenodo.22679573`.

41. Stellbrink, L. & Nunner, H. hnunner/pg-literature-screening: v1.0.0 – Pandemic Guidelines literature screening (2026). `https://zenodo.org/doi/10.5281/zenodo.22647106`.

Table 1: **The 40 consensus recommendations.** Statements appear in the wording the panel rated, which keeps US spelling. Must act is our judgement, not the panel's: M modelling groups; A public health agencies, including funders and providers of national data infrastructure; P policy advisors (Supplementary Note 10). Ratings run from 1 (completely disagree) to 6 (completely agree), and all 40 met the consensus criterion of IQR ≤ 1. The top-two share is the percentage of raters choosing 5 or 6, which still separates statements once the median saturates. FAIR4RS, in B02, extends the FAIR principles to research software. n varies because partial responses were retained and panellists could answer "Not qualified to respond" on any statement. Supplementary Tables 4 and 5 give the Round 3 comparison and a completers-only reanalysis.

ID	Recommendation	Must act	Median	IQR	Top-two share (%)	n
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.	M	6	0	98	51
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.	M	6	1	98	51

continued on next page

Table 1: *(continued)*

ID	Recommendation	Must act	Median	IQR	Top-two share (%)	n
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.	M	6	0	98	50
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.	M	6	1	92	50
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.	M	6	0	96	50
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.	M	6	1	90	50

continued on next page

Table 1: *(continued)*

ID	Recommendation	Must act	Median	IQR	Top-two share (%)	n
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.	M	6	1	96	49
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.	M	6	1	94	49
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.	M	6	1	94	49
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.	M	5.5	1	92	50

continued on next page

Table 1: *(continued)*

ID	Recommendation	Must act	Median	IQR	Top-two share (%)	n
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.	M	5	1	80	45
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.	M, A	6	1	86	49
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.	M	6	0	92	49
Section B: Data availability and quality						

continued on next page

Table 1: *(continued)*

ID	Recommendation	Must act	Median	IQR	Top-two share (%)	n
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.	M, A	6	0	96	49
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.	M, A	6	1	90	49
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.	A	6	1	90	49
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.	A	6	1	90	48

continued on next page

Table 1: *(continued)*

ID	Recommendation	Must act	Median	IQR	Top-two share (%)	n
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.	A	6	1	90	48
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.	A	6	1	87	47
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.	A	6	1	92	48
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.	A, P	6	0	93	46
Section C: Uncertainty, validation, and communication						

continued on next page

Table 1: *(continued)*

ID	Recommendation	Must act	Median	IQR	Top-two share (%)	n
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.	M	6	1	94	47
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.	M	6	0	96	47
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.	M	6	0.75	98	46

continued on next page

Table 1: *(continued)*

ID	Recommendation	Must act	Median	IQR	Top-two share (%)	n
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.	M, A	6	1	98	43
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.	M	6	1	94	48
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.	M, A	6	1	85	48

continued on next page

Table 1: *(continued)*

ID	Recommendation	Must act	Median	IQR	Top-two share (%)	n
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.	A	6	0	98	46
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.	A	6	0	92	48
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.	M, P	6	1	77	48

continued on next page

Table 1: *(continued)*

ID	Recommendation	Must act	Median	IQR	Top-two share (%)	n
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.	M, A	6	1	90	48
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.	A	6	0	96	47
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.	A	6	0	100	47
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.	A, P	6	0	98	47

continued on next page

Table 1: *(continued)*

ID	Recommendation	Must act	Median	IQR	Top-two share (%)	n
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.	A	6	1	87	47
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.	M, A	6	1	91	44
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.	M	6	0	96	45

continued on next page

Table 1: *(continued)*

ID	Recommendation	Must act	Median	IQR	Top-two share (%)	n
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.	M	6	1	85	47
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.	A	6	1	84	45
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.	A	6	0.5	96	47

Table 2: **Review themes mapped onto the Delphi statements.** The themes were synthesised after the statement set was fixed, so this maps coverage retrospectively rather than how statements were generated. Percentages are the share of the 135 publications addressing each theme (Supplementary Table 3) and order the rows; the term-based codebook behind them (Supplementary Table 10) agrees only moderately with hand labels (Cohen's $\kappa = 0.55$, recall 0.62), so they rank themes rather than measure them. Themes are not mutually exclusive, so A13, C05, C09, and C10 each appear twice. The rows are our reading of which statements bear on each theme, not the codebook applied to statement wording: the codebook was built for publication text, and run over the statements it would both miss and add rows. A statement is listed only where we judged its subject to fall under a theme, which leaves 21 of the 40 in no row at all: the Section A items bearing on model construction (A02, A04, A06 to A09, A11) are absent from methodological improvement, and A01, A03, A05, B03, B05 to B07, C03, C04, C06, D04 to D06, and D09 match no theme. Coverage and Delphi weight diverge: *Model development and methodological improvement* is the most discussed yet carries two statements, while *Data governance, sharing, and infrastructure*, *Uncertainty quantification and communication*, and *Capacity building and global equity* carry four each. The final column links to Table 3, and each row names at most one priority, so four of the ten appear in no row: external model validation systems, model-selection guidance, protocols for multi-model ensembles, and rapid ethical governance frameworks, each resting mainly on statements that match no theme.

Review theme (% of publications)	**Statements covering it**	**Round 4 consensus**	**Linked priority**
Model development and methodological improvement (60%)	A10, B04	All, but A10 among the lowest-rated (median 5.5)	Behavioural and social science integration
Data governance, sharing, and infrastructure (50%)	B01, C07, C08, D03	All, and all four at the ceiling (median 6, IQR 0)	Interoperable real-time surveillance
Uncertainty quantification and communication (47%)	C01, C02, C09, C10	All	Standardised uncertainty communication
Science-policy interface (36%)	A12, D01	All	Pre-established science-policy structures
Capacity building and global equity (33%)	B08, D02, D07, D08	All, but their top-two shares range from 84% (D08) to 100% (D02)	Global cooperation and equity-sensitive models
Public and stakeholder communication (27%)	C09, C10	All	No distinct priority, absorbed into standardised uncertainty communication
Model transparency and reproducibility (23%)	A13, B02, C05	All	Transparent, reproducible model reporting
Reporting standards and documentation (10%)	A13, C05	All	No distinct priority, absorbed into transparent, reproducible model reporting

Table 3: **Ten priorities for the inter-pandemic period.** Priorities synthesised from the consensus recommendations and the review, grouped by what each demands and listed alphabetically within each group. The panel rated neither the priorities nor their relative importance, so no order here implies any: agreement is not importance, since a statement can command consensus yet matter little for preparedness. The panel did rank seven preparedness measures across three phases, and those do not correspond one to one with these ten (Supplementary Table 7). "Must act" uses M, A, and P as in Table 1, and is the union of the parties named for the supporting statements. It is our judgement rather than the panel's. Every priority requires modelling groups to act, but only two of the ten can be delivered by a modelling group alone. A statement can support more than one priority: C02 and B08 each appear in two rows.

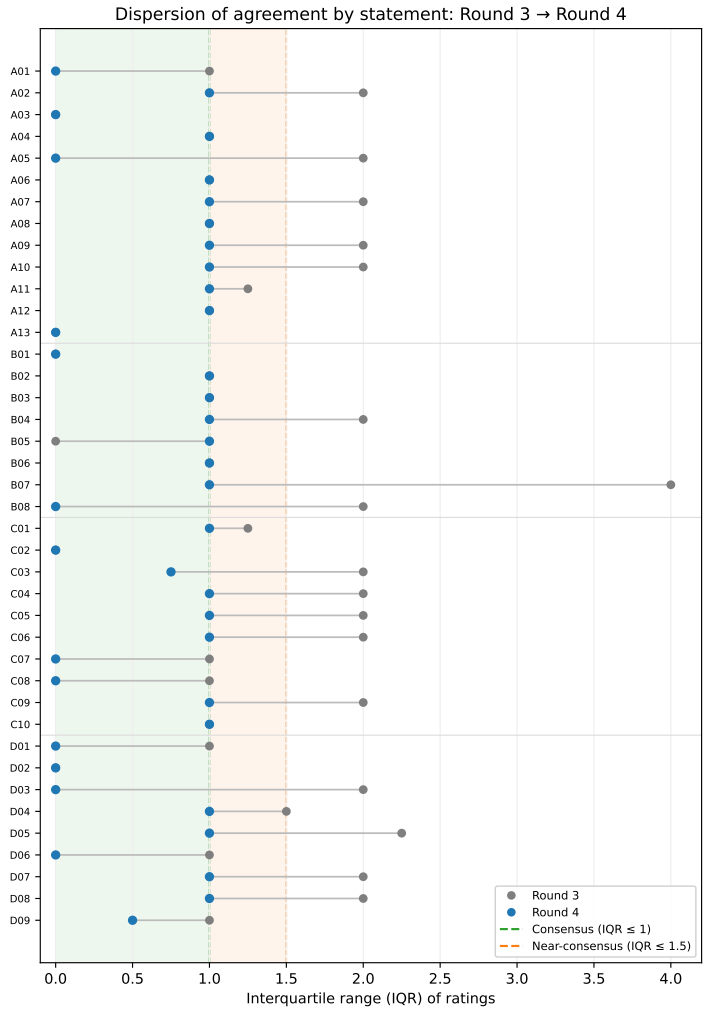
Priority	What it asks for	Must act	Supporting consensus recommendations (topic)
Data and infrastructure			
External model validation systems	Test performance on data not used for fitting, keep validation distinct from calibration, and update it as empirical data arrive.	M, A	C04 (validation)
Interoperable real-time surveillance	Timely data with documented limitations, released in machine-readable form through free, documented, versioned interfaces, and cross-border sharing agreements settled in advance.	M, A, P	B01, C07, C08, D03 (data governance and infrastructure)
Modelling practice and methods			
Behavioural and social science integration	Justify socio-economic and behavioural detail against the question asked, and collect socio-behavioural data alongside clinical surveillance from the outset.	M, A	A10, B04 (model development and methodological improvement)
Model-selection guidance (fitness for purpose)	Judge a model against the question it addresses, and state the simplifying assumptions and boundary conditions a simple model rests on.	M	A01, A04 (model design and complexity)

continued on next page

Table 3: *(continued)*

Priority	What it asks for	Must act	Supporting consensus recommendations (topic)
Protocols for multi-model ensembles	Combine structurally different models so that convergence signals robustness and divergence exposes structural uncertainty, then quantify that uncertainty by type.	M	A03, C02 (quantification of structural uncertainty)
Standardised uncertainty communication	Report uncertainty as ranges rather than point values, tie it to the decisions a model informs, and convey it to policymakers, the public, and the media alike.	M, A, P	C01, C02, C09, C10 (uncertainty quantification and communication)
Transparent, reproducible model reporting	Report calibration method, parameters, sources, and assumptions; share models, data, and code under the FAIR principles; and attach sensitivity analyses to every policy-informing output.	M, A	A13, B02, C05 (transparency and reproducibility)
Governance, cooperation, and ethics			
Global cooperation and equity-sensitive models	Joint modelling hubs and capacity building in under-resourced settings, age-stratified impacts on vulnerable groups, community advisory mechanisms, and ethical frameworks for data inequity.	M, A, P	B08, D02, D07, D08 (capacity building and global equity)
Pre-established science-policy structures	Standing interdisciplinary collaborations, and models scoped to the decision context and the data actually available.	M, A	A12, D01 (science-policy interface)
Rapid ethical governance frameworks	Embed ethics upstream in model development, validation, and communication, identify and mitigate bias, and agree safeguards before a pandemic rather than during one.	M, A, P	B08, D05, D06 (collaboration, interdisciplinarity, and ethics)

Extended Data



Extended Data Fig. 1: **Cross-round dispersion of agreement in Round 3 and Round 4.** Points give the interquartile range (IQR) of the six-point ratings for each of the 40 statements in Round 3 (grey) and Round 4 (blue), connected by a line and grouped by section. Round 3 rated 42 statements; merging two pairs gave the 40 of Round 4, and each merged item's Round 3 figures come from the pooled distributions of its two statements (Supplementary Note 6 and Supplementary Table 6). Shaded bands mark the consensus (IQR ≤ 1) and near-consensus ($1 <$ IQR ≤ 1.5) zones. Every statement lies in the consensus zone at Round 4, and dispersion falls or holds for all but one (B05, 0 to 1, still within the zone).

Extended Data Table 1: **Composition of the expert panel.** Characteristics of the 41 panellists who completed the final Round 4 rating. Panellists selected from pre-defined response options for every item. The career-stage categories use one common set of academic titles, which do not map identically onto every national system. Self-rated expertise is the panellist's own assessment rather than an external judgement. Background, sector, and career stage are separate questions, so their counts need not agree: two panellists work outside a university while one reports a non-academic background. Percentages are of these 41 completers and may not sum to 100 due to rounding.

Characteristic (self-reported)	n	%
Geographic region of primary affiliation		
Europe	36	88
North America	2	5
Latin America and the Caribbean	1	2
Oceania	2	5
Professional background		
Academic / researcher	40	98
Clinician / practitioner	1	2
Employment sector		
University / research institute	39	95
Hospital / health system	1	2
Independent practice / self-employed	1	2
Career stage (current position)		
Full professor / senior researcher	27	66
Associate professor / senior lecturer	8	20
Assistant professor / junior faculty	3	7
Doctoral researcher	2	5
Non-academic professional	1	2
Years of professional experience in the field		
More than 20 years	14	34
11 to 20 years	13	32
5 to 10 years	9	22
Less than 5 years	5	12
Self-rated expertise		
Recognised expert / leadership	12	29
Extensive / active involvement	12	29
Substantial experience	11	27
Some / limited depth	6	15
Published on the topic within the past 5 years		
Yes	40	98
No	1	2

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- PreparingfortheNextPandemicTakesMoreThanBetterModelsSupplementaryInformation.pdf