PLOS Computational Biology

RESEARCH ARTICLE

Calibration of transmission-dynamic infectious disease models: A scoping review and reporting framework

Emmanuelle A. Dankwa[1]*, Léa Cavalli[2], Ruchita Balasubramanian[2], Melike Hazal Can[1], Hening Cui[1], Katherine M. Jia[2], Yunfei Li[1], Sylvia K. Ofori[2], Nicole A. Swartwood[1], Carrie G. Wade[3], Caroline O. Buckee[4], Jeffrey W. Imai-Eaton[2,5], Nicolas A. Menzies[1]

1 Department of Global Health and Population, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, United States of America, **2** Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America, **3** Countway Library, Harvard School of Medicine, Boston, Massachusetts, United States of America, **4** Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, United States of America, **5** MRC Centre for Global Infectious Disease Analysis, School of Public Health, Imperial College London, London, United Kingdom

* edankwa@hsph.harvard.edu

OPEN ACCESS

Citation: Dankwa EA, Cavalli L, Balasubramanian R, Can MH, Cui H, Jia KM, et al. (2025) Calibration of transmission-dynamic infectious disease models: A scoping review and reporting framework. PLoS Comput Biol 21(11): e1013647. https://doi.org/10.1371/journal.pcbi.1013647

Editor: Joseph T. Wu, University of Hong Kong, HONG KONG

Received: March 13, 2025

Accepted: October 22, 2025

Published: November 4, 2025

Data availability statement: All data relevant to this study are included or linked in the article. The code and relevant data dependencies for reproducing the results in this study are available at the repository: https://github.com/Leacavalli/Calibration-Review.

Abstract

Objective/Background

Transmission-dynamic models are commonly used to study infectious disease epidemiology. Calibration involves identifying model parameter values that align model outputs with observed data or other evidence. Inaccurate calibration and inconsistent reporting produce inference errors and limit reproducibility, compromising confidence in the validity of modeled results. No standardized framework exists for reporting on calibration of infectious disease models, and an understanding of current calibration approaches is lacking.

Methods

We developed the Purpose-Inputs-Process-Outputs (PIPO) framework for reporting calibration practices and applied it in a scoping review to assess calibration approaches and evaluate reporting comprehensiveness in transmission-dynamic models of tuberculosis, HIV and malaria published between January 1, 2018, and January 16, 2024. We searched relevant databases and websites to identify eligible publications, including peer-reviewed studies where these models were calibrated to empirical data or published estimates.

Results

We identified 411 eligible studies encompassing 419 models, with 74% (n = 309) being compartmental models and 20% (n = 81) individual-based models (IBMs). The

Funding: This project was funded by the Centers for Disease Control and Prevention (200-2016-91779 to COB and NAM). The findings, conclusions, and views expressed are those of the author(s) and do not necessarily represent the official position of the Centers for Disease Control and Prevention (CDC). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

predominant analytical purpose was to evaluate interventions (71% of models, n = 298). Parameters were calibrated mainly because they were unknown or ambiguous (40%, n = 168), or because determining their value was relevant to the scientific question beyond being necessary to run the model (20%, n = 85). The choice of calibration method was significantly associated with model structure (p-value<0.001) and stochasticity (p-value = 0.006), with approximate Bayesian computation more frequently used with IBMs and Markov-Chain Monte Carlo with compartmental models. Regarding reporting comprehensiveness, all PIPO framework items were reported in 4% (n = 18) of models; 11–14 items in 66% (n = 277), and 10 or fewer items in 28% (n = 124). Implementation code was the least reported, available in only 20% (n = 82) of models.

Conclusions

Reporting on calibration is heterogeneous in recent infectious disease modeling literature. Our proposed framework for reporting of calibration approaches could support improved reproducibility and credibility of modeled analyses.

Author summary

Calibration, the identification of parameter values so that model outcomes are consistent with observed data or other evidence, is often employed in the process of obtaining model results to inform health decision making. Despite its importance, there has not been a standardized framework for reporting how calibration is conducted in infectious disease modeling studies. This has led to inconsistent reporting practices and challenges in reproducing model results, potentially compromising confidence in their validity. We developed a calibration reporting framework, based on best practices found in the literature and informed by our expertise in conducting calibration. To assess calibration practices and their reporting, we applied our framework in a scoping review of 419 infectious disease transmission models of HIV, TB and malaria published between 2018 and 2024. Most models reviewed were compartmental (74%) or individual-based (20%), and the choice of calibration methods was associated with model structure and stochasticity. Calibration was conducted predominantly in the context of models aimed at evaluating the impact of disease control interventions, highlighting the role of calibration in decision making. Parameters were calibrated mainly because they were unknown or ambiguous, or because reporting their value was relevant to the scientific question beyond just being necessary to run the model. The comprehensiveness of calibration reporting varied across models, with most models omitting 1–5 items in the framework. Accessible implementation code was the most underreported, with only 20% of models including it. Our proposed framework could serve as a tool to standardize calibration reporting, thereby enhancing the transparency and reproducibility of calibration processes in transmission-dynamic models.

1 Introduction

Infectious disease models are designed to reproduce key features of disease epidemiology, including the impact of health services or interventions on the transmission or morbidity processes. These models are often used to predict disease trends or evaluate alternative interventions for disease control. Models are characterized by parameters: fixed values or variables that determine model behavior. Model calibration (or 'model fitting') includes a diverse range of methods for selecting values for model parameters such that the model yields estimates consistent with existing evidence. Specifically, a calibrated model is one in which the value of *at least one* parameter is chosen to achieve consistency of model outputs with empirical data, published estimates, or other evidence. In practice, this is commonly achieved by applying formal numerical optimization or statistical approaches that systematically vary parameters and assesses them according to a quantitative goodness-of-fit measure.

Calibration of infectious disease transmission models may be undertaken to infer the value of a parameter of epidemiological importance. These parameter estimates are interpreted as defining characteristics of the modeled diseases' epidemiology, for example, the duration of the latent or incubation periods. Calibration can also be employed to enable the prediction of disease trends under a range of interventions. Such predictions are used as evidence to support policies on disease control; which attained widespread public prominence for the control of the COVID-19 pandemic [1,2].

Inaccuracies in model calibration may result in inference errors, compromising the validity of modeled results that inform public health policies. Calibration is one among several choices that influence the validity of modeled evidence, such as model structure [3], methodological assumptions [3] and data type, all of which impact parameter identifiability [4–7]. The clarity, and therefore credibility, of model results may also be adversely influenced by an inadequate description of the calibration procedure employed in a study. This lack of thorough description and non-reporting of implementation code hampers reproducibility [8], and potentially compromises trust in the validity of those studies for informing public health action.

Although calibration approaches are widely employed in infectious disease modeling, few studies have detailed their application, with most literature consisting of tutorials or best practices guidelines [9–14]. In addition, efforts to develop a standardized framework for calibration have been limited. The Infectious Disease Modeling Reproducibility Checklist (IDMRC) [8] provides a general framework for reporting modeling study components to ensure reproducibility but overlooks details specific to calibration, such as the uncertainty of calibration outputs [15] or the number of parameter sets calibrated. [11] proposed an extension to [16] checklist for calibration reporting based on a review of individual-based transmission models (IBMs), but it has limited applicability to other model structures, like compartmental models. A standardized framework for calibration in infectious disease modeling could enhance objective and consistent reporting, ensuring that reports include all relevant details necessary for reproducibility. Moreover, no comprehensive overview exists of calibration approaches across all model types, examining how study context influences the choice of approach and evaluating reporting comprehensiveness. Understanding current calibration practices and their reporting could improve method selection and reporting and help identify areas for innovation and further research.

To address these gaps, we developed the *Purpose-Input-Process-Output (PIPO)* framework for reporting infectious disease model calibration. We applied this framework in a scoping review of the literature on transmission-dynamic models for tuberculosis (TB), HIV, and malaria published between January 1, 2018, and January 16, 2024. The review systematically mapped the conduct and reporting comprehensiveness of calibration in this field. It focused on evaluating the purpose, inputs, process, and outputs of calibration in recent literature, and examined how the choice of calibration approach varies by model type.

2 Methods

2.1 Purpose-input-process-output (PIPO) framework

We developed the *Purpose-Input-Process-Output (PIPO) framework* as a proposed framework and checklist for reporting on infectious disease model calibration. PIPO is a 16-item reporting framework for describing calibration in infectious disease modeling studies to ensure reproducibility of calibration by facilitating a clear communication of calibration aims, methods and

results. This framework was developed based on the authors' expertise in conducting calibration for transmission-dynamic models, and published guidance on calibration best practices [9–14]. The framework has four broad components: 1) *Purpose*, which deals with the goal of calibration, 2) *Inputs*, which deal with the inputs into the calibration algorithm, 3) *Process*, which deals with how calibration is conducted, and 4) *Outputs*, which deals with the characteristics of the calibration outputs.

A brief overview of each component of the framework is as follows.

Purpose: This component collects information on the scientific problem being addressed by the study. For example, is the goal to understand disease mechanisms or to predict disease trends? Understanding the goal helps establish the context for the calibration.

Inputs: This component allows the reporting of key characteristics of the main inputs for calibration: 1) the parameters to be calibrated, and 2) the calibration targets. For parameters, PIPO includes fields to report on whether prior knowledge on any parameters is incorporated into the calibration process, whether all or just a subset of parameters was calibrated, and to justify why these parameters were selected for calibration. We use "prior knowledge" to refer to any pre-existent data, estimate or expert opinion about a parameter that is used to inform parameter calibration. For calibration targets, PIPO allows reporting on the number of calibration targets, the type of data used for defining targets (e.g., incidence, spatial data, etc.) and whether the targets are empirical data (as raw counts or numbers and their corresponding statistical summaries, such as means, medians, rates, or proportions), or modeled estimates. The absence of specific information on parameters and calibration targets when reproducing a calibration result has many implications. For example, parameter estimates obtained through calibration may differ depending on which parameters are fixed (held at a constant value in the model) and which are calibrated. Also, the type of data used as a calibration target could affect calibrated parameter estimates. For instance, the identifiability of parameters may vary depending on whether incidence, prevalence or another type of data is used as calibration target [4]. Therefore, without clear reporting on the characteristics of calibration targets and parameters, the reproducibility of calibration and model results is hampered.

Process: This component allows the reporting of all details related to the calibration method used. Reporting covers the name of the calibration method and a brief description of how it works, including its goodness-of-fit (GOF) measure(s), which assesses the level of agreement between modeled outcomes and calibration targets. Further, the component allows reporting the following details of the calibration method's implementation: 1) a well-documented, accessible repository for calibration code (if used), 2) programming language, and 3) versions for any programming languages, packages, or data repository used. In our review of models, we examined the availability of any model implementation code, acknowledging that calibration code is not typically reported separately.

Output: Characteristics of calibrated parameter estimates and the corresponding model outcomes, jointly termed "calibration outputs" are reported in this component. Reporting items include specification of whether the parameter values produced by calibration represented a "point estimate" (i.e., a single parameter or parameter set), a "sample estimate" (i.e., multiple parameter values or sets) or a "distribution estimate" (i.e., a closed-form distribution function that could be used to generate new parameter values), as can be obtained with Laplace approximation [17]. Details on the size of calibration outputs and uncertainty in the output are also reported here, as they are relevant for reproducibility. For calibration processes that produce a sample of parameter values or sets, an insufficient sample size can introduce additional uncertainty into modelled outcomes. For this reason, the sample size of calibrated parameters used for inferences must be clearly reported. Finally, any model validation methods are reported here.

The reporting framework is in S1 Text and has been formatted such that it is readily available for use as a checklist for calibration reporting.

2.2 Scoping review

2.2.1 Review conduct and reporting. We reported the review following the Preferred Reporting Items for Systematic Review and Meta-analysis Extension for Scoping Reviews (PRISMA-ScR) checklist [18] (S2 Table). In developing the

protocol and conducting the review, we followed the guidelines proposed by [19]. The review protocol was pre-registered with the Open Science Framework (https://doi.org/10.17605/OSF.IO/3VTJW).

2.2.2 Eligibility criteria and information sources. We used the Studies, Data, Methods and Outcomes (SDMO) Framework [20] to define the inclusion and exclusion criteria based on the research question: In transmission-dynamic models of TB, HIV and malaria, how is calibration conducted? Inclusion and exclusion criteria are detailed in S1 Table and summarized as follows: Published and peer-reviewed studies in which a transmission-dynamic model for HIV, TB or malaria was calibrated to empirical data or published estimates were included. Following [21], we defined a dynamic infectious disease transmission model as a mathematical model for describing infection or disease spread, where the risk of infection is not constant and depends on the number of infectious individuals in the population at any given time, thus allowing for the modeling of nonlinear feedback effects. We focused on dynamic models, acknowledging that the norms of calibration of dynamic models may differ from those of other models, such as static models, which make the assumption that the risk of infection is constant over time [22]. We defined calibration as the use of a method to select values or distributions for model parameters so that model outputs were consistent with observed data or estimates, referred to as *calibration targets*. Unpublished studies, studies from proceedings, opinion pieces or animal studies were excluded. We also excluded studies that were not published in the English language due to limitations with translating and interpreting foreign language studies. To focus our review on current practices, we restricted our search to studies published within approximately 5 years of the search date, covering the period 1 January 2018 to 16 January 2024 (the search date). S3 Table describes the full search strategy, which was constructed with librarian support (CW). Briefly, we searched PubMed, Embase, lobal Health, Web of Science Core Collection, and Global Index Medicus databases to identify eligible studies that contain in their title or abstract keywords related to “transmission-dynamic models”, and “HIV”, “tuberculosis” or “malaria”. We also searched relevant websites of major disease consortia for HIV (HIV Modeling Consortium: http://hivmodeling.org/) and TB (TB Modelling and Analysis Consortium: https://tb-mac.org/). Although we identified a consortium for malaria vaccine models, we did not add this to our search because we sought for collections that were general and not topic specific.

2.2.3 Selection of sources of evidence.

2.2.3.1 Title and abstract screening: Before title and abstract screening, the screening review team (EAD, LC, RB, MHC, KMJ and SKO) pilot tested the inclusion/exclusion criteria as follows: A random sample of 14 titles/abstracts were selected and each team member independently screened these using the eligibility criteria. An agreement of 85% was obtained. The team then discussed the discrepancies and modified the criteria as needed. For title/abstract screening, we applied all inclusion and exclusion criteria except the criterion that studies needed to involve calibration. This is because calibration may not always be mentioned in the abstract even if it was performed in a study. Each title and abstract were reviewed independently by two reviewers, with conflicts adjudicated by a third reviewer.

2.2.3.2 Full text screening: Studies not screened out based on their titles and abstracts underwent full text screening. Most full texts were either freely available online or accessible via institutional library access. We excluded studies for which full texts were not accessible. Each full text was read independently by two reviewers, and the full inclusion and exclusion criteria were applied. Conflicts were resolved by a third reviewer.

2.2.4 Data extraction. We extracted studies using the PIPO framework. The data extraction form for this review was expanded to additionally include general study characteristics such as the year of publication, disease of focus and model structure. The form had 22 reporting items in total: 15 items corresponding to the first 15 items in the PIPO framework and 7 items on general study characteristics. We note that extraction was done based on a first version of the framework, which we subsequently updated to reflect additional items later identified as worth reporting. The form was developed with input from all authors and tested by a team of four reviewers (EAD, LC, RB, KMJ). After incorporating feedback from the first round of testing, two reviewers (EAD and LC) conducted a second instance of testing. The final data extraction form is in S4 Sheet. For each study, data were independently extracted by two reviewers in the data extraction team (EAD, LC, RB, MHC, YL, NS, HC). Conflicts were resolved by a third reviewer in the team.

2.2.5 Synthesis of results.

2.2.5.1 Evaluate the conduct of calibration in the recent literature: First, we summarized the extracted data as counts and percentages for every extracted item. Additionally, we performed statistical tests to assess the relationship between choice of calibration method, model structure and model stochasticity. To ensure large sample sizes and sufficient statistical power, statistical tests were limited to calibration methods that were used in at least 30 models and to model structure or stochasticity categories that had at least one report for each of these calibration methods. For the test involving model stochasticity, we employed a chi-square test. For the test involving model structure, we used the Fisher's exact test as expected values were small for some cells hence the Chi-square approximation, which assumes large samples, could not be used. Logistic regression was used to determine the relationship between calibration methods and model structure and stochasticity. The figures of test results were produced using the "ggstatsplot" package (v 0.13.0) [23] in R (v 4.3.1).

2.2.5.2 Evaluate the comprehensiveness of calibration reporting: To investigate the extent to which calibration reporting was comprehensive, we defined a 15-item comprehensiveness scale, corresponding to the first 15 items in the PIPO reporting framework. For each study, we assigned a score of 1 if an item was reported or a score of 0 if it was not reported. Therefore, the maximum attainable score was 15, indicating a study with high reporting comprehensiveness, and the minimum attainable score was 0, indicating a study with low reporting comprehensiveness. We categorized studies by score as follows. A study with a score of 0–5 was categorized as having "low" reporting comprehensiveness; a study with a score of 6–10 was categorized as having "fair" reporting comprehensiveness, and a study with a score of 11–15 was categorized as having "high" reporting comprehensiveness. We interpreted reporting comprehensiveness as indicative of transparency around calibration methods and potential for reproducibility.

We performed a series of statistical tests to ascertain whether significant differences exist in code reporting and overall reporting comprehensiveness based on journal or publisher. We restricted the analysis to publishers and journals for which there were more than 5 models. We used a Fisher's exact test to investigate the association between journal or publisher and code reporting while we used a Kruskal-Wallis test to investigate the association between journal or publisher and overall reporting comprehensiveness score. In addition to these univariate analyses, we used two logistic regression models for bivariate analysis which included both journal and publisher with code reporting as outcome in one model and overall reporting comprehensiveness as outcome in the other. We also assessed the correlation between models' reporting score and the publishing journal's 5-year impact factor.

2.2.6 Analysis tools. Screening and data extraction were performed in Covidence (Veritas Health [24]). Analyses of extracted data were performed in R version 4.3.1. [25].

2.2.7 Reproducibility. The full, commented analysis code and relevant data dependencies are available at the repository: https://github.com/Leacavalli/Calibration-Review.

2.2.8 Results validation. One reviewer (LC) conducted data cleaning on the raw extracted data to address inconsistencies due to varying reporting styles across studies and reviewers. This included standardizing free-text entries that were reported differently. The data cleaning code is available on GitHub. In addition, when inconsistencies in the extracted data were suspected during analysis, two reviewers (LC and EAD) checked the relevant sections of the extracted data and where necessary, also checked with the original data extractor to ensure accuracy.

3 Results

3.1 Study selection

The literature search yielded 6518 studies. After duplicates were removed, 3138 studies remained to be screened by title and abstract. Of these, 765 progressed to the full text review stage. We excluded 354 of these studies based on the exclusion criteria or because they did not have accessible full texts (n = 4). We included 411 studies in the review (Fig 1).

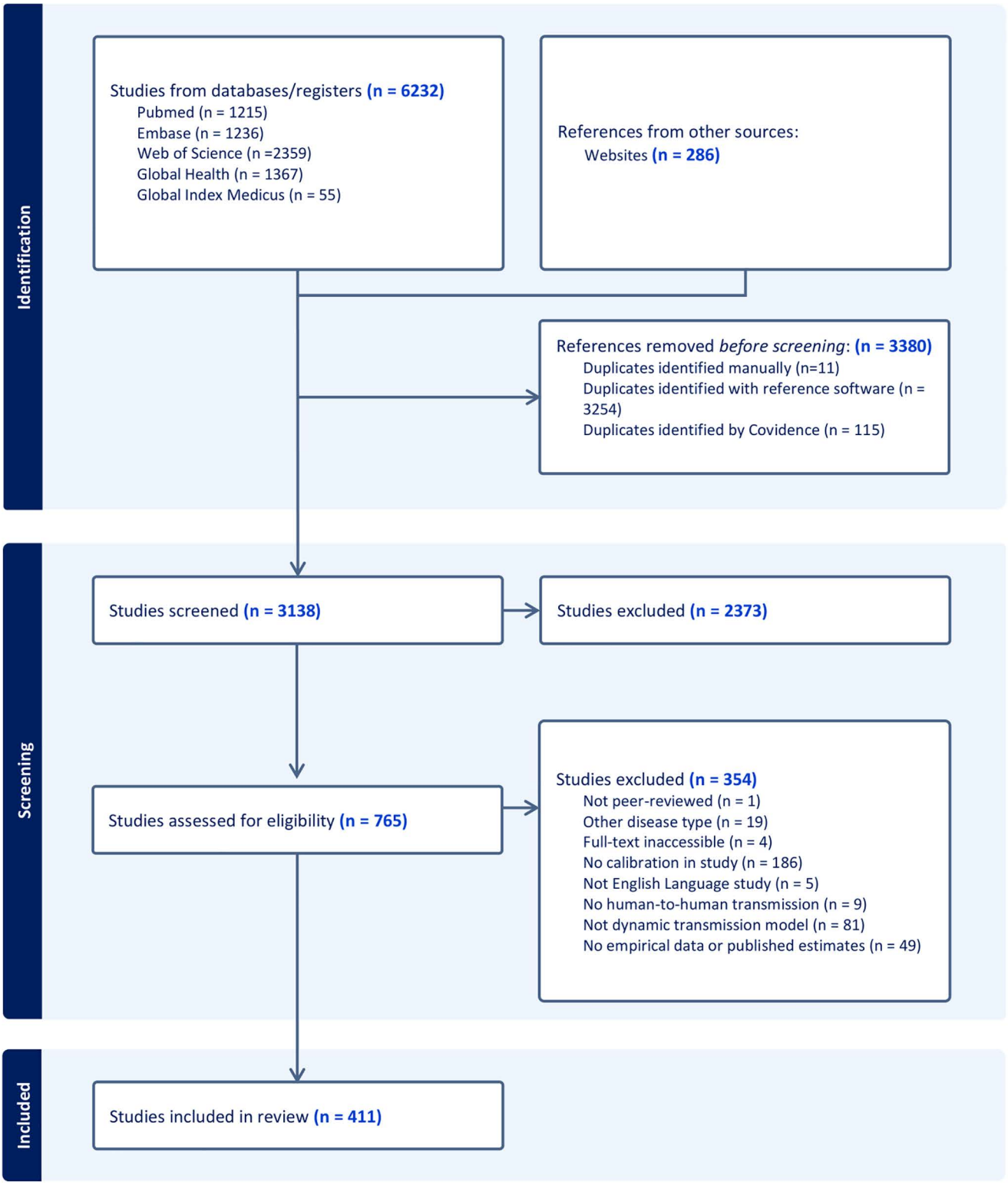

Fig 1. PRISMA flow chart of study selection process.

https://doi.org/10.1371/journal.pcbi.1013647.g001

3.2 General model characteristics

The 411 included studies yielded 419 unique model applications and their calibrations, as five studies calibrated multiple models (S4 Table). The full extracted data for each calibrated model are in S1 Sheet. Summary statistics (i.e., counts and percentages, where applicable) for each extracted item are in S2 Sheet.

Table 1 summarizes the general characteristics of all 419 calibrated models. Of the models studied, 48% (n = 203) focused exclusively on HIV, 33% (n = 138) on TB, 16% (n = 67) on malaria, and 3% (n = 11) on combined HIV/TB. In structure, models were mostly compartmental (74%, n = 309) or individual-based (20%, n = 81). One model [26] used a Hawkes process. Model structure was unreported in 6% (n = 23) of the models. Deterministic models accounted for 36% (n = 149) of reviewed models, while stochastic models accounted for 22% (n = 91). Whether or not a model included stochasticity was unreported in the remaining 43% (n = 179).

3.3 Conduct of calibration in the literature

3.3.1 Purpose. Predominantly, calibration was conducted in the context of evaluating or comparing interventions (71% of models, n = 298). Other purposes were to understand disease mechanisms (38%, n = 161) or predict disease trends (24%, n = 102). The least frequent purpose of calibration was to assess the impact of model assumptions (9%, n = 37). Some models (40%, n = 169) were calibrated for multiple reasons (therefore, percentages add to >100%).

3.3.2 Inputs. For most models (63%, n = 264), prior knowledge about parameters to be calibrated was incorporated in the calibration process. In many cases, this prior knowledge was obtained from the literature, evidenced by citations of other papers for parameter sources. In most models (92%, n = 384), calibration was limited to a subset of model parameters. Only 12 models (3%), where reported, had all parameters calibrated. Parameters were selected for calibration because they were unknown or ambiguous (40%, n = 168), and/or because determining and reporting their

Table 1. General characteristics of calibrated models (n = 419).

Characteristic	Categories	Count	Percentage (n = 419)
Year published	2018	73	17.4
	2019	60	14.3
	2020	76	18.1
	2021	77	18.4
	2022	71	16.9
	2023	60	14.3
	2024	2	0.5
Disease type	HIV only	203	48.4
	Malaria only	67	16.0
	TB only	138	33.0
	HIV/TB	11	2.6
Model structure	Compartmental	309	73.7
	Individual-based	81	19.6
	Hawkes Process	1	0.2
	Not reported	23	5.5
	Other	5	1
Stochasticity in model	Deterministic model	149	35.6
	Stochastic model	91	21.7
	Not reported	179	42.7

https://doi.org/10.1371/journal.pcbi.1013647.t001

accurate value was relevant to the question of interest beyond just being necessary to run the model (20%, n = 85). Most model reports (45%, n = 190) did not include a justification for the choice of parameters to be calibrated.

The main types of data used for defining calibration targets were disease prevalence (48% of models, n = 201), notifications or diagnoses (46%, n = 193), incidence (42%, n = 176), treatment-related data (39%, n = 165) and demographic data (28%, n = 117) (S1 Fig). The least used data types for calibration were spatial data (n = 3), cost data (n = 1) and effect sizes from trials (n = 1). Calibration targets were mostly empirical data or their corresponding statistical summaries (86%, n = 359), while the use of modelled estimates was less common (34%, n = 142). Most models were calibrated to multiple targets (72%, n = 300) rather than a single calibration target (26%, n = 110). In 2% (n = 9) of models, the number of calibration targets was not reported.

3.3.3 Process.

3.3.3.1 Calibration methods: The four most frequently used calibration methods were least squares estimation (20%, n = 83), Markov Chain Monte Carlo (MCMC) methods (16%, n = 68), Approximate Bayesian Computation (ABC) (10%, n = 40) and maximum likelihood estimation (8%, n = 32). Each of these methods were employed in at least 30 models. We classified studies under Sequential Monte Carlo (SMC) if they reported their calibration method as "Monte-Carlo filtering" or "SMC based on particle filtering", both of which are interchangeably used in the literature to refer to SMC. Studies that reported using ABC-SMC were classified under ABC because ABC-SMC primarily uses an ABC framework but with SMC integration to improve proposals of parameter values [27].

In 20 models (5%), parameters were calibrated manually without relying on a numerical optimisation algorithm, i.e., by hand-tuning.

We classified calibration methods into two broad families according to the aim of the calibration procedure, inferred from the nature of the calibration results as extracted from a study (see item D.1. in the PIPO framework). If the nature of the calibration result was a "Sample estimate" or "Distribution estimate", we concluded that the aim of the procedure was to approximate a distribution. If the nature of the calibration result was a "Point estimate", we concluded that the aim of the procedure was to identify a single optimal parameter set. For studies for which the nature of the calibration result was not clear or judged as "Other", we revisited the study for more details to determine the most appropriate aim of the calibration procedure. We reviewed 18 methods aiming to identify an optimal parameter set and 7 methods aiming to approximate a distribution (Table 2). A brief description for each calibration method we reviewed is provided in S5 Table.

The choice of calibration method was significantly associated with type of model structure (p-value < 0.001, S2 Fig) and model stochasticity (p-value = 0.009, S3 Fig). The ABC method was used significantly more frequently with IBMs than with compartmental models (odds ratio (OR)= 7.7, 95% CI: 3.4-17.8, p < 0.001), with 50% of IBMs (n = 17/34) and 12% of compartmental models (n = 20/175) being calibrated using ABC (S2 Fig). Conversely, MCMC methods were used significantly more frequently with compartmental models than with IBMs (OR=3.6, 95% CI: 1.3-12.6, p = 0.021), with 12% of IBMs (n = 4/34) and 33% of compartmental models (n = 57/175) being calibrated using MCMC (S3 Fig). The odds of using least square and maximum likelihood estimation did not significantly differ by model structure (respectively: OR=0.5, 95% CI: 0.2-1.1, p = 0.10, and OR= 0.8, 95% CI: 0.2-2.2, p = 0.70). The association patterns of calibration methods with model stochasticity were similar, with the ABC method being significantly more common with stochastic models (OR=4.1, 95% CI: 1.7-10.0, p = 0.001), while the MCMC method was more frequently used with deterministic models, although this relationship was not statistically significant (OR=2.1, 95% CI: 0.9-5.1, p = 0.078). The least square and maximum likelihood estimation methods showed no significant association with stochasticity (least square: OR=0.6, 95% CI: 0.3-1.3, p = 0.21; maximum likelihood estimation: OR=1.0, 95% CI: 0.4-2.8, p = 0.95). This aligns with the tight link between model structure and stochasticity, since 88% of compartmental models with reported stochasticity were deterministic (n = 80/91), while 97% of IBMs (n = 30/31) were stochastic.

3.3.3.2 Goodness-of-fit (GOF) measures: The GOF measures employed as part of the calibration process were based on an ad-hoc distance function (27%, n = 115), data likelihood (22%, n = 93) or another measure (5%, n = 22), such

Table 2. Calibration methods classified by aim of calibration procedure. The number and percentage of models applying these methods have also been indicated. Brief descriptions of methods are in S5 Table.

Calibration method	Number of models	Percentage of models (n = 419)
Aim: **Approximate a distribution**		
Markov Chain Monte Carlo (MCMC)	68	16.2
Approximate Bayesian Computation (ABC)	40	9.5
Incremental Mixture Importance Sampling (IMIS)	15	3.6
Sequential Monte Carlo (SMC)	11	2.6
Sampling importance resampling (SIR)	9	2.1
History matching with emulation (HME)	7	1.7
Laplace Approximation	1	0.2
Aim: **Identify an optimal parameter set**		
Least squares estimation	83	19.8
Maximum likelihood estimation	32	7.6
Manual (hand-tuning)	20	4.8
Nelder-Mead algorithm	6	1.4
Grey Wolf Optimizer (GWO)	3	0.7
Genetic algorithm	2	0.5
Parallel Simultaneous Perturbation Optimisation (PSPO)	2	0.5
Broyden-Fletcher-Goldfarb-Shanno algorithm	2	0.5
Coordinate descent algorithm	1	0.2
Deviance-based loss	1	0.2
Levenberg-Marquardt optimization algorithm	1	0.2
Local sequential quadratic programming (DESQP) optimization algorithm	1	0.2
Maximum-a-Posteriori estimation	1	0.2
Mean square error with threshold	1	0.2
Method of moments	1	0.2
Minimum Step Deviation	1	0.2
Minimum contrast estimation	1	0.2
Root Mean Squared (RMS) deviation	1	0.2

https://doi.org/10.1371/journal.pcbi.1013647.t002

as the deviance information criterion [28], and the Akaike information criterion [29–31]. For 45% of models (n = 189), the GOF measure was either not reported at all or not reported sufficiently clearly to allow extraction of relevant information. Although we did not specifically assess the predictive power of models, one of the studies [32] used cross-validation, a method for assessing predictive power, as a goodness-of-fit measure. Specifically, they used area under the curve (AUC) scores from leave-one-out cross validation as a measure of model fit.

3.3.3.3 Software: Calibration implementation code was reported in an open-access repository for only 20% of models (n = 82). For 9 models (2%), the calibration code was inaccessible although a link to a repository was reported. For most models (78%, n = 328), calibration code was not reported at all. Regarding programming languages used for calibration, the following were the most used: R (26%, n = 110), MATLAB (20%, n = 84), C++ (8%, n = 33) and Python (6%, n = 23), in order of frequency of use. Notably, for 181 models (43%, n = 181), the programming language used was not reported. Multiple programming languages were used in 33 models (8%).

3.3.4 Output. Where reported, calibration outputs were typically point estimates (42%, n = 174) or a "sample estimate" (i.e., multiple parameter values or sets, 45%, n = 189). Only one model [17] presented results as a "distribution

estimate". In this model, calibration was achieved through a Laplace approximation process, which provided a closed-form approximation of the posterior distributions for both parameter values and model outputs. Reporting of calibration outputs was exclusively numerical (10%, n = 40), exclusively graphical (13%, n = 53) or, in most cases (74%, n = 311), a combination of numerical and graphical approaches. Calibration outputs were not reported at all in 15 models (4%). Regarding uncertainty in parameter estimates, we observed both numerical (i.e., as confidence or credible intervals) and graphical (i.e., shaded areas around curves on graphs) reporting in 41% (n = 170) of models, exclusive numerical reporting in 12% (n = 52) and exclusive graphical reporting in 15% (n = 64). Many model reports (32%, n = 133) did not report on the uncertainty in their parameter estimates.

Among calibration processes that generated a multiple sets of parameter values (i.e., a "sample estimate"), there was substantial variation in the size of calibration outputs (S4 Fig). The largest calibration outputs, with over 2000 parameter sets, were only observed with MCMC, ABC, and Incremental Mixture Importance Sampling.

3.3.5 Comprehensiveness of calibration reporting. The least reported item in the PIPO framework was the implementation code, which was reported and accessible in only 20% of models (n = 82).

Other less reported items—reported in fewer than 70% of models—were the justification for the choice of parameters to calibrate (55%, n = 229), whether calibration was done in a single step or sequentially (56%, n = 234), the GOF measure used within the calibration process (57%, n = 238), whether external beliefs or evidence was used for calibration (66%, n = 276) and the uncertainty in parameter estimates (68%, n = 286).

Items with high reporting frequencies were the type (99%, n = 416) and resolution (95%, n = 397) of data used for defining calibration targets, the number of calibration targets (98%, n = 410), calibration outputs (96%, n = 404) and the choice of parameters to calibrate (95%, n = 396).

The reporting comprehensiveness was excellent (all 15 items reported) in 4% of models (n = 18), high (11–14 items reported) in 66% of models (n = 277), fair (6–10 items reported) in 28% (n = 118) and low (0–5 items reported) in 1% (n = 6) (Fig 2). We did not observe differences in the distributions of reporting comprehensiveness by year of model publication (S5 Fig). Individual reporting item scores for all 419 models are presented in S3 Sheet.

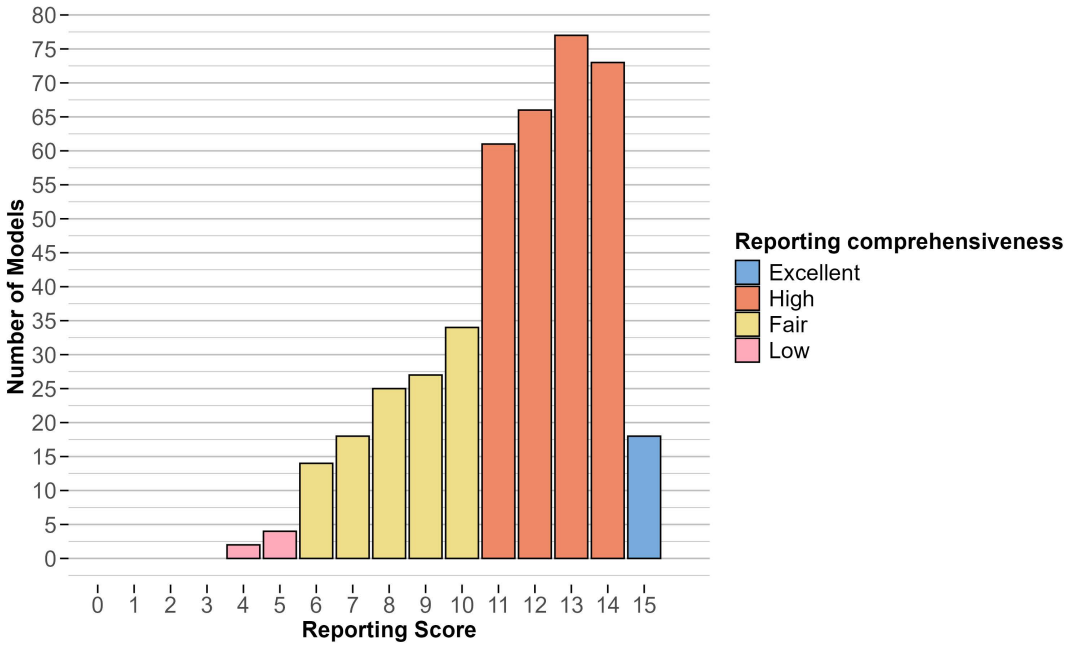

Fig 2. Distribution of scores for calibration reporting comprehensiveness for 419 models.

https://doi.org/10.1371/journal.pcbi.1013647.g002

We found differences in code availability to be significantly associated with both the journal in which a model was published (Fisher's exact test, p = 0.0005), and the publisher (Fisher's exact test, p = 0.0025, S6 Fig) in univariate analyses. In a logistic regression model including both journal and publisher, only journal remained a significant predictor of code availability (p < 0.001), while publisher did not. We found differences in reporting scores to be significantly associated with journal (chi-squared = 35.29, df = 19, p-value = 0.013, S7 Fig) but not with publisher (chi-squared = 10.526, df = 10, p-value = 0.396, S8 Fig). However, individual models' reporting score showed a weak correlation with the publishing journal's 5-year impact factor (Pearson's r = 0.15, p = 0.027).

4 Discussion

We developed a framework for calibration reporting, encompassing criteria regarding the purpose, inputs, process, and outputs (PIPO) of calibration, based on best practices on calibration in the literature, and the authors' expertise in conducting calibration. Our PIPO framework addresses the shortcomings of previous reporting frameworks [8,11] in several ways. Specifically, PIPO is applicable to all model structures, whereas the framework in [11] was primarily designed for IBMs. Additionally, PIPO includes a field for code reporting, which is crucial for ensuring reproducibility. We recommend using the PIPO framework as a complement to the IDMRC, as it specifically expands on elements specific to calibration reporting not captured in the IDMRC, such as the uncertainty and size of calibration outputs, which are essential for model reproducibility.

We applied the PIPO framework to examine current calibration conduct and reporting practices through a scoping review of 419 models from 411 studies on HIV, TB, and malaria transmission-dynamic modelling, published between 2018 and 2024. Our review showed that in 30% of models (n = 124), fewer than 11 items on the PIPO framework were reported, with the lowest reporting observed for calibration implementation code, reported in only 20% of models (n = 82). This is lower than was observed in a recent review of early COVID-19 modeling studies [33]. We found that variation in code availability and overall reporting comprehensiveness was associated with journal rather than by publisher. Although code availability guidelines are available at the publisher level, they may only be effective to the extent that individual journals enforce them. Efforts to promote efficient code reporting, such as enforcing data and code reporting requirements by journals, the development of tools and best practice guidelines [34] and including reproducible software development in the training of modelers [35] should be encouraged.

We observed that a large fraction of model reports (45%, n = 190) did not justify the selection of specific parameters for calibration over others. Explicitly considering whether a parameter needs to be calibrated prompts important questions including whether calibration targets are appropriate for informing the parameter's value/distribution, whether alternative data on the parameter exists, and whether calibration will be relevant for addressing the modelling problem. When considered earlier in the study process, such questions could promote efficiency by reducing time spent exploring calibration avenues that may not be feasible given the study design and available data. In line with the findings of [11], many models in our review (45%, n = 189) did not report or were unclear about the GOF measure used. Given that the GOF measure evaluates the level of agreement between modeled outcomes and calibration targets, it is essential to define and report the GOF measure clearly and quantitatively, rather than relying on subjective or non-quantitative measures, thereby ensuring reproducibility and credibility of studies.

Most models (86%, n = 359) relied on empirical data or their summaries — examples including prevalence and notifications — for defining calibration targets (S1 Fig). Because of this reliance on empirical data, efforts to improve its quality and completeness should be emphasized. While not captured by the PIPO framework, we encourage authors to evaluate the quality of empirical data prior to its use, as it could directly influence the quality of parameter estimates. Efforts to improve data collection systems and the quality of collected data will not only aid in infectious disease surveillance, but also in the development of more accurately calibrated models. Furthermore, while not evaluated in our review, potential errors or missingness in the data also needs to be accounted for or corrected prior

to data use for calibration to minimize biases in estimates. When modeled estimates rather than empirical data are used to define calibration targets, it is important to consider how the underlying uncertainty in these targets could be reflected in the uncertainty in parameter estimates. For some of the models we reviewed (68%, n = 286), uncertainty estimates for their calibration outputs were reported but we did not study the sources of this uncertainty. A potential area of future work is to explore how uncertainty in calibration targets, as well as other types of uncertainty (model structure uncertainty, stochastic uncertainty) are captured in uncertainty estimates for parameters. We recommend the clear reporting of uncertainty, as this is useful for promoting reproducibility and transparent communication with public health decision-makers.

The choice of calibration method varied with the type of model structure, based on statistical tests of independence. The ABC method, for example, was used much more frequently with IBMs than with compartmental models (S2 Fig). These differences can be explained in part by the compatibility between model structure and calibration method. Compared to compartmental models, IBMs are more complex in structure and therefore tend to have more complex likelihood functions that are difficult to specify. Given that ABC allows for likelihood-free inference [36], it is suitable for use with IBMs for which a likelihood function may be unavailable. The significant relationship between model stochasticity and calibration method could similarly be explained, at least in part, by model complexity. For instance, complex IBMs, usually associated with likelihood-free calibration methods like ABC (S2 Fig), are also typically stochastic [37]. This co-occurrence may explain why, among stochastic models, we observed a greater preference for methods such as ABC and a lower preference for likelihood-dependent methods such as MCMC (S3 Fig).

This review has a few limitations. First, the calibration practices examined here may have limited generalizability beyond models with structures like those used for HIV, malaria and TB and identified in the review. These diseases were selected because they are among the most extensively studied epidemics, contributing significantly to the global burden of disease. They also encompass a diverse range of transmission routes and natural history patterns, and thus a variety of model types. Nevertheless, their coverage is not exhaustive. Alternative transmission routes, such as the fecal-oral route exhibited by cholera and other gastrointestinal illness, are missing. Given that transmission routes and natural history influence model structure, we suspect that some model types are not represented. Second, there is a potential language bias, as the review was limited to studies published in the English language, such that we may have missed model types that only tend to be published in other languages. Third, we did not track repeated uses of the same core model, which could have potentially biased our results to reflect characteristics of models that are more generalizable or adaptable. Finally, although all items in the PIPO framework are important for reproducibility, one could argue that differences may exist in the relative importance of these items. We could not assess these differences in our work. We recommend this as a direction for future work.

In conclusion, we have evaluated the conduct and reporting of calibration in 419 recently published infectious disease transmission models. In this review, the conduct and reporting of calibration was found to be highly heterogeneous, indicating a potential role for a standardized reporting framework to enhance reproducibility of results and consequently, the credibility of model results used for health decision making. Based on information from the reviewed models, best practices in literature and authors' expertise in calibration, we developed the PIPO calibration reporting framework that captures four important components of calibration: its purpose, inputs, process and outputs. Use of this framework as a reporting tool could enhance standards in calibration reporting and improve the credibility and reproducibility of infectious disease model results. Moving forward, it is essential to continually monitor potential shift in the types of methods and adapt calibration methods and reporting standards accordingly.

Supporting information

S1 Table. Inclusion and exclusion criteria based on the Studies, Data, Methods and Outcome (SDMO) framework.
(DOCX)

S2 Table. Preferred Reporting Items for Systematic reviews and Meta-Analyses extension for Scoping Reviews (PRISMA-ScR) Checklist. From Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMAScR): Checklist and Explanation. Ann Intern Med. 2018;169:467–473. https://doi.org/10.7326/M18-0850.
(DOCX)

S3 Table. Full electronic search* strategy.
(DOCX)

S4 Table. Details of studies with multiple calibrated models.
(DOCX)

S5 Table. Detailed description of calibration methods.
(DOCX)

S1 Fig. Frequency of data types used for defining calibration targets.
(TIFF)

S2 Fig. Most frequent calibration methods (used in at least 30 models) and associated model structure.
(TIFF)

S3 Fig. Most frequent calibration methods (used in at least 30 models) and associated model stochasticity.
(TIFF)

S4 Fig. Distribution of the size of calibration output for outputs greater than size 1. Results are shown for calibration methods reported in at least 10 models.
(TIFF)

S5 Fig. Distribution of reporting comprehensiveness by year of model publication.
(TIFF)

S6 Fig. Percentage of models (y-axis) in a journal (x-axis) for which code was reported. n = number of models per journal in study. Journals are ordered by publisher on the x-axis. Figure is limited to journals with more than five models in the study.
(TIFF)

S7 Fig. Reporting scores by journal. n = number of models per journal in study. Navy numeric annotations indicate the median score per journal while firebrick annotations indicate the mean score.
(TIFF)

S8 Fig. Reporting score by publisher. n = number of models per publisher in study. Navy numeric annotations indicate the median score per publisher while firebrick annotations indicate the mean score.
(TIFF)

S1 Sheet. Full extracted data for each calibrated model.
(XLSX)

S2 Sheet. Summary statistics (i.e., counts and percentages, where applicable) for each extracted item.
(XLSX)

S3 Sheet. Individual reporting item scores for all models.
(XLSX)

S4 Sheet. Data extraction form.
(XLSX)

S1 Text. Purpose-Inputs-Process-Outputs (PIPO) calibration reporting framework.
(DOCX)

Acknowledgments

We would like to thank Jesse Knight (Imperial College London) for his valuable inputs on the calibration reporting framework.

Author contributions

Conceptualization: Emmanuelle A Dankwa.

Data curation: Emmanuelle A Dankwa, Léa Cavalli.

Formal analysis: Emmanuelle A Dankwa, Léa Cavalli.

Funding acquisition: Caroline O Buckee, Nicolas A Menzies.

Investigation: Emmanuelle A Dankwa, Léa Cavalli, Ruchita Balasubramanian, Melike Hazal Can, Hening Cui, Katherine M Jia, Yunfei Li, Sylvia K Ofori, Nicole A Swartwood, Carrie G. Wade.

Methodology: Emmanuelle A Dankwa, Caroline O Buckee, Jeffrey W Imai-Eaton, Nicolas A Menzies.

Supervision: Nicolas A Menzies.

Visualization: Emmanuelle A Dankwa.

Writing – original draft: Emmanuelle A Dankwa.

Writing – review & editing: Léa Cavalli, Ruchita Balasubramanian, Melike Hazal Can, Hening Cui, Katherine M Jia, Yunfei Li, Sylvia K Ofori, Nicole A Swartwood, Carrie G. Wade, Caroline O Buckee, Jeffrey W Imai-Eaton, Nicolas A Menzies.

References

1. Brooks-Pollock E, Danon L, Jombart T, Pellis L. Modelling that shaped the early COVID-19 pandemic response in the UK. Philos Trans R Soc Lond B Biol Sci. 2021;376(1829):20210001. https://doi.org/10.1098/rstb.2021.0001 PMID: 34053252
2. Centers for Disease Control and Prevention. COVID-19 Forecasting and Mathematical Modeling [WWW Document]. Centers for Disease Control and Prevention. 2020. Accessed 2023 December 15. URL https://www.cdc.gov/coronavirus/2019-ncov/science/forecasting/forecasting-math-modeling.html
3. Brisson M, Edmunds WJ. Impact of model, methodological, and parameter uncertainty in the economic analysis of vaccination programs. Med Decis Making. 2006;26(5):434–46. https://doi.org/10.1177/0272989X06290485 PMID: 16997923
4. Dankwa EA, Brouwer AF, Donnelly CA. Structural identifiability of compartmental models for infectious disease transmission is influenced by data type. Epidemics. 2022;41:100643. https://doi.org/10.1016/j.epidem.2022.100643 PMID: 36308994
5. Eisenberg MC, Robertson SL, Tien JH. Identifiability and estimation of multiple transmission pathways in cholera and waterborne disease. J Theor Biol. 2013;324:84–102. https://doi.org/10.1016/j.jtbi.2012.12.021 PMID: 23333764
6. Kao Y-H, Eisenberg MC. Practical unidentifiability of a simple vector-borne disease model: Implications for parameter estimation and intervention assessment. Epidemics. 2018;25:89–100. https://doi.org/10.1016/j.epidem.2018.05.010 PMID: 29903539
7. Tuncer N, Le TT. Structural and practical identifiability analysis of outbreak models. Math Biosci. 2018;299:1–18. https://doi.org/10.1016/j.mbs.2018.02.004 PMID: 29477671
8. Pokutnaya D, Childers B, Arcury-Quandt AE, Hochheiser H, Van Panhuis WG. An implementation framework to improve the transparency and reproducibility of computational models of infectious diseases. PLoS Comput Biol. 2023;19(3):e1010856. https://doi.org/10.1371/journal.pcbi.1010856 PMID: 36928042
9. Briggs AH, Weinstein MC, Fenwick EAL, Karnon J, Sculpher MJ, Paltiel AD, et al. Model parameter estimation and uncertainty: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--6. Value Health. 2012;15(6):835–42. https://doi.org/10.1016/j.jval.2012.04.014 PMID: 22999133

10. Chowell G. Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A Primer for parameter uncertainty, identifiability, and forecasts. Infect Dis Model. 2017;2(3):379–98. https://doi.org/10.1016/j.idm.2017.08.001 PMID: 29250607
11. Hazelbag CM, Dushoff J, Dominic EM, Mthombothi ZE, Delva W. Calibration of individual-based models to epidemiological data: A systematic review. PLoS Comput Biol. 2020;16(5):e1007893. https://doi.org/10.1371/journal.pcbi.1007893 PMID: 32392252
12. Jackson CH, Jit M, Sharples LD, De Angelis D. Calibration of complex models through Bayesian evidence synthesis: a demonstration and tutorial. Med Decis Making. 2015;35(2):148–61. https://doi.org/10.1177/0272989X13493143 PMID: 23886677
13. Menzies NA, Soeteman DI, Pandya A, Kim JJ. Bayesian Methods for Calibrating Health Policy Models: A Tutorial. Pharmacoeconomics. 2017;35(6):613–24. https://doi.org/10.1007/s40273-017-0494-4 PMID: 28247184
14. Vanni T, Karnon J, Madan J, White RG, Edmunds WJ, Foss AM, et al. Calibrating models in economic evaluation: a seven-step approach. Pharmacoeconomics. 2011;29(1):35–49. https://doi.org/10.2165/11584600-000000000-00000 PMID: 21142277
15. Ryckman T, Luby S, Owens DK, Bendavid E, Goldhaber-Fiebert JD. Methods for Model Calibration under High Uncertainty: Modeling Cholera in Bangladesh. Med Decis Making. 2020;40(5):693–709. https://doi.org/10.1177/0272989X20938683 PMID: 32639859
16. Stout NK, Knudsen AB, Kong CY, McMahon PM, Gazelle GS. Calibration methods used in cancer simulation models and suggested reporting guidelines. Pharmacoeconomics. 2009;27(7):533–45. https://doi.org/10.2165/11314830-000000000-00000 PMID: 19663525
17. Maheu-Giroux M, Marsh K, Doyle CM, Godin A, Lanièce Delaunay C, Johnson LF, et al. National HIV testing and diagnosis coverage in sub-Saharan Africa: a new modeling tool for estimating the "first 90" from program and survey data. AIDS. 2019;33 Suppl 3(Suppl 3):S255–69. https://doi.org/10.1097/QAD.0000000000002386 PMID: 31764066
18. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med. 2018;169(7):467–73. https://doi.org/10.7326/M18-0850 PMID: 30178033
19. Peters M, Godfrey C, McInerney P, Munn Z, Tricco A, Khalil H. Chapter 11: Scoping reviews (2020 version). In: Aromataris E, Munn Z. (Eds.), JBI Manual for Evidence Synthesis. JBI; 2020. https://doi.org/10.46658/jbimes-20-12
20. Munn Z, Stern C, Aromataris E, Lockwood C, Jordan Z. What kind of systematic review should I conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences. BMC Med Res Methodol. 2018;18(1):5. https://doi.org/10.1186/s12874-017-0468-4 PMID: 29316881
21. Pitman R, Fisman D, Zaric GS, Postma M, Kretzschmar M, Edmunds J, et al. Dynamic transmission modeling: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--5. Value Health. 2012;15(6):828–34. https://doi.org/10.1016/j.jval.2012.06.011 PMID: 22999132
22. Porgo TV, Norris SL, Salanti G, Johnson LF, Simpson JA, Low N, et al. The use of mathematical modeling studies for evidence synthesis and guideline development: A glossary. Res Synth Methods. 2019;10(1):125–33. https://doi.org/10.1002/jrsm.1333 PMID: 30508309
23. Patil I. Visualizations with statistical details: The "ggstatsplot" approach. JOSS. 2021;6(61):3167. https://doi.org/10.21105/joss.03167
24. Veritas Health Innovation. Covidence systematic review software. 2024.
25. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2024.
26. Unwin HJT, Routledge I, Flaxman S, Rizoiu M-A, Lai S, Cohen J, et al. Using Hawkes Processes to model imported and local malaria cases in near-elimination settings. PLoS Comput Biol. 2021;17(4):e1008830. https://doi.org/10.1371/journal.pcbi.1008830 PMID: 33793564
27. Sisson SA, Fan Y, Tanaka MM. Sequential Monte Carlo without likelihoods. Proc Natl Acad Sci U S A. 2007;104(6):1760–5. https://doi.org/10.1073/pnas.0607208104 PMID: 17264216
28. Shaweno D, Trauer JM, Denholm JT, McBryde ES. The role of geospatial hotspots in the spatial spread of tuberculosis in rural Ethiopia: a mathematical model. R Soc Open Sci. 2018;5(9):180887. https://doi.org/10.1098/rsos.180887 PMID: 30839742
29. Kim S, Byun JH, Park A, Jung IH. A mathematical model for assessing the effectiveness of controlling relapse in Plasmodium vivax malaria endemic in the Republic of Korea. PLoS One. 2020;15(1):e0227919. https://doi.org/10.1371/journal.pone.0227919 PMID: 31978085
30. Mettle FO, Osei Affi P, Twumasi C. Modelling the Transmission Dynamics of Tuberculosis in the Ashanti Region of Ghana. Interdiscip Perspect Infect Dis. 2020;2020:4513854. https://doi.org/10.1155/2020/4513854 PMID: 32318105
31. Mussina K, Kadyrov S, Kashkynbayev A, Yerdessov S, Zhakhina G, Sakko Y, et al. Prevalence of HIV in Kazakhstan 2010-2020 and Its Forecasting for the Next 10 Years. HIV AIDS (Auckl). 2023;15:387–97. https://doi.org/10.2147/HIV.S413876 PMID: 37426767
32. Routledge I, Lai S, Battle KE, Ghani AC, Gomez-Rodriguez M, Gustafson KB, et al. Tracking progress towards malaria elimination in China: Individual-level estimates of transmission and its spatiotemporal variation using a diffusion network approach. PLoS Comput Biol. 2020;16(3):e1007707. https://doi.org/10.1371/journal.pcbi.1007707 PMID: 32203520
33. Pokutnaya D, Van Panhuis WG, Childers B, Hawkins MS, Arcury-Quandt AE, Matlack M, et al. Inter-rater reliability of the infectious disease modeling reproducibility checklist (IDMRC) as applied to COVID-19 computational modeling research. BMC Infect Dis. 2023;23(1):733. https://doi.org/10.1186/s12879-023-08729-4 PMID: 37891462
34. Piccolo SR, Frampton MB. Tools and techniques for computational reproducibility. Gigascience. 2016;5(1):30. https://doi.org/10.1186/s13742-016-0135-4 PMID: 27401684

35. Gallagher K, Creswell R, Lambert B, Robinson M, Lok Lei C, Mirams GR, et al. Ten simple rules for training scientists to make better software. PLoS Comput Biol. 2024;20(9):e1012410. https://doi.org/10.1371/journal.pcbi.1012410 PMID: 39264985

36. Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. Genetics. 2002;162(4):2025–35. https://doi.org/10.1093/genetics/162.4.2025 PMID: 12524368

37. Willem L, Verelst F, Bilcke J, Hens N, Beutels P. Lessons from a decade of individual-based models for infectious disease transmission: a systematic review (2006-2015). BMC Infect Dis. 2017;17(1):612. https://doi.org/10.1186/s12879-017-2699-8 PMID: 28893198