Calibration of individual-based models to epidemiological data: a systematic review

C. Marijn Hazelbag[1*], Jonathan Dushoff[2], Emanuel M. Dominic[1], Zinhle E. Mthombothi[1], Wim Delva[1,3,4,5,6,7]

[1] The South African Department of Science and Technology-National Research Foundation (DST-NRF) South African Centre for Epidemiological Modelling and Analysis (SACEMA), Stellenbosch University, Stellenbosch, South Africa

[2] Department of Mathematics and Statistics, the Institute for Infectious Disease Research, McMaster University, Hamilton, Ontario, Canada

[3] School for Data Science and Computational Thinking, Stellenbosch University, Stellenbosch, South Africa

[4] Center for Statistics, I-BioStat, Hasselt University, Diepenbeek, Belgium

[5] Department of Global Health, Faculty of Medicine and Health, Stellenbosch University, Stellenbosch, South Africa

[6] International Centre for Reproductive Health, Ghent University, Ghent, Belgium

[7] Rega Institute for Medical Research, KU Leuven, Leuven, Belgium

* Corresponding author

E-mail: marijnhazelbag@sun.ac.za

Abstract

Individual-based models (IBMs) informing public health policy should be calibrated to data and provide estimates of uncertainty. Two main components of model-calibration methods are the parameter-search strategy and the goodness-of-fit (GOF) measure; many options exist for each of these. This review provides an overview of the calibration methods used in IBMs modelling infectious disease spread.

We identified articles on PubMed employing simulation-based methods to calibrate IBMs in HIV, tuberculosis, and malaria epidemiology that were published between 2013 and 2018. Articles were included if models stored individual-specific information, and calibration involved comparing model output to population-level targets. We extracted information on parameter-search strategies and GOF measures. We also recorded the reporting of model validation.

The PubMed search identified 653 candidate articles, of which 84 met the criteria for review. Of the included articles, 40 (48%) combined a quantitative measure of GOF with an algorithmic parameter-search strategy, which was an optimisation algorithm (14 articles) or a sampling algorithm (26 articles). These 40 articles varied widely in their choices of parameter-search strategies and GOF measures. For the remaining 44 articles, the parameter-search strategy could either not be identified (32 articles) or was described as an informal, non-reproducible method (12 articles). Of these 44 articles, the majority (25 articles) were unclear about the GOF measure used; of the rest, only five quantitatively evaluated GOF. Thirty-two (38%) articles reported on model validation.

Less than half of the articles reviewed used algorithmic parameter-search strategies. Only one-third of articles quantified the uncertainty around calibrated parameter values and could incorporate this uncertainty into their model projections. There was no consensus on which algorithmic calibration method to use. Only 38% of the articles performed model validation. The

adoption of better-documented algorithmic calibration and validation methods could improve both the reproducibility and the quality of inference in model-based epidemiology.

Introduction

Individual-based models (IBMs) intended to inform public health policy should be calibrated to real-world data and provide valid estimates of uncertainty [1,2]. IBMs track information for a simulated collection of interacting individuals [3]. IBMs allow for more detailed incorporation of heterogeneity, spatial structure, and individual-level adaptation (e.g. physiological or behavioural changes) compared to other modelling frameworks [4]. This makes IBMs valuable planning tools, particularly in settings where real-world complexities that are not accounted for in simpler models have important effects [5,6]. However, researchers and policymakers often battle with the question of how much value they can attach to the results of IBMs [7]. Fitting an IBM to empirical data (calibration) improves confidence that the simulation model provides a realistic and accurate estimate of the outcome of health policy decisions (e.g. projection of the disease prevalence under different intervention strategies, or the cost-effectiveness of different intervention strategies) [8-12]. Transparent reporting on calibration methods for IBMs is therefore required [11,12].

Parameter values with accompanying confidence intervals used in IBMs are obtained from the literature and are often obtained through statistical estimation. When researchers cannot estimate parameters from empirical data, they obtain their likely values through calibration [12]. Parameter calibration is often difficult for IBMs because their greater complexity renders exact calculation of the likelihood impossible [13,14]. Consequently, simulation-based calibration methods that approximate the likelihood have been developed [15]. These consist of running the model for different parameter sets to identify parameter sets producing model output that best resembles the summary statistics obtained from the empirical data (e.g. disease prevalence over time). To do this, formal simulation-based calibration requires *summary statistics* (*targets*) from empirical data, a *parameter-search strategy* for exploring the parameter space, a *goodness-of-fit (GOF)* measure to evaluate the concordance between model output and targets, *acceptance criteria* to determine which parameter sets produce model

output close enough to the targets, and a *stopping rule* to determine when the calibrations ends [9,16].

In this review, we pay particular attention to the parameter-search strategy and GOF measure used. Algorithmic parameter-search strategies can be divided into *optimisation algorithms* and *sampling algorithms* [14]. Optimisation algorithms find the parameter combination that optimises the GOF, resulting in a single best parameter combination. Examples include grid-search and simplex-based/direct search methods (e.g. the Nelder-Mead method) [17], but many different algorithms exist [18]. A grid search typically does not provide a confidence interval around parameter estimates, whereas different options for obtaining confidence intervals exist for other algorithms (e.g. the profile likelihood method, Fisher information, etc.) [19,20]. Sampling algorithms aim to find a distribution of parameter values that approximate the likelihood surface or posterior distribution. Examples include approximate Bayesian computation (ABC) methods and sampling importance resampling [8,13,14,21,22]. Parameter distributions obtained from sampling algorithms allow for the representation of correlations between parameters and for parameter uncertainty to be incorporated into model projections [2,6,8,16,23]. Quantitative measures of GOF include squared distances, relative distances, and approximations of the likelihood. A more subjective method of calibration involves the manual adjustment of parameter values, followed by a visual assessment of whether the model outputs resemble empirical data [24].

Previous research in the context of IBMs of HIV transmission found that 22 (69%) out of 32 articles described the process through which the model was calibrated to data [12]. The impact of stochasticity on model results was summarized in nearly half (n=15) of the articles [12]. The depth of reporting on calibration methods was highly variable [9,12]. Previous research on calibration methods in cancer-simulation models in general – not IBMs specifically – found that 131/154 (85%) articles may have calibrated at least one unknown parameter. Of the 131 articles that calibrated parameters, 84 (64%) did not describe the use of a GOF measure, 27

(21%) used a quantitative GOF such as the likelihood or distance measures, and the remaining 20 (15%) used visual assessment of GOF [9]. Only a few articles reported parameter distributions resulting from calibration; most only presented a single best parameter combination [9]. Information on the parameter-search strategy and stopping rules was generally not well described, and acceptance criteria were rarely mentioned [9,25]. Of the 154 articles included in the review by Stout *et al.,* 80 (52%) mentioned model validation [9]. However, while previous studies have reviewed specific portions of the modelling literature, they either did not focus on IBMs or did not focus on the calibration methods in much detail.

We conducted a systematic review of epidemiological studies using IBMs of the HIV, malaria and tuberculosis (TB) epidemics, as these have been among the most investigated epidemics. We aim to provide an overview of current practices in the simulation-based calibration of IBMs.

Materials and methods

This review was performed following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement [26]. The PRISMA flow diagram details the selection process of articles included for review (see Figure 1).

Search strategy and selection criteria

We identified articles on PubMed that employed simulation-based methods to calibrate IBMs of HIV, malaria and tuberculosis, and that were published between 1 January 2013 and 31 December 2018. The following search query was performed on 31 January 2019: *'((HIV[tiab] OR malaria[tiab] OR tuberculo*[tiab] OR TB[tiab]) AND (infect* OR transmi* OR prevent*) AND (computer simulation[tiab] OR microsimulation[tiab] OR simulation[tiab] OR agent-based[tiab] OR individual-based[tiab] OR computer model*[tiab] OR computerized model*[tiab]) AND ("2013/01/01"[Date - publication] : "2018/12/31"[Date - publication]) NOT(molecular))'*.

Eligibility criteria were agreed upon by WD, JD and CMH before screening. Articles were included if models stored individual-specific information and calibration involved running the model and comparing model output to population-level targets expressed as summary statistics. We excluded review articles, statistical simulation studies, and studies that focused on molecular biology and immunology because we were primarily interested in studies informing public health policy.

Titles and abstracts were screened for eligibility by CMH, and difficult cases were discussed with WD. If the title and abstract did not provide sufficient information for exclusion, a full-text examination was performed. Full-text inclusion was performed by two independent researchers (CMH and either ZM or ED) for a subset of 100 articles. CMH included 28 articles, of which six were not included by ZM and ED; these six articles were double-checked by WD and consequently included for review. ZM included four articles that CMH did not include; these four articles were double-checked by WD and consequently not included for review. After that, full-text inclusion was performed by CMH in consultation with WD.

Data extraction

For each article, we extracted information on the goal, parameter-search strategy and the GOF measure, the rationale for choosing this calibration strategy over another, and model validation. Acceptance criteria and stopping rules are only relevant for articles applying algorithmic parameter-search strategies and collected for that subset of articles. For readability purposes, we say "used" to mean "reported the use of" throughout this review.

To ensure the reliability of the collection of information on calibration, this information was collected independently by two reviewers (CMH and either ZM or ED) for each article included using a prospectively developed form. This form was based on the Calibration Reporting Checklist of Stout *et al.* [9] and was extended by several items. Reviewers received

specific instructions for finding sections on calibration. Information on calibration methods was extracted verbatim, allowing for later classification. Articles on which there was disagreement in the classification were discussed by WD, JD and CMH until an agreement was reached. We classified articles reporting both algorithmic and informal calibration as informal since doing part of the calibration informally makes the entire calibration irreproducible.

We set out to collect information on the number of calibrated parameters, the number of fixed parameters, and the number of targets. However, we observed large interrater differences, indicating the difficulty of this task. We, therefore, report only ranges for these counts.

Results

Selection of articles for inclusion

The PubMed search resulted in 653 publications, of which 84 articles were included for review; 388 were excluded based on title and abstract, and another 181 were excluded based on a full-text review (see Fig 1). The number of articles selected by publication year increased from seven in 2013 to 20 in 2018.

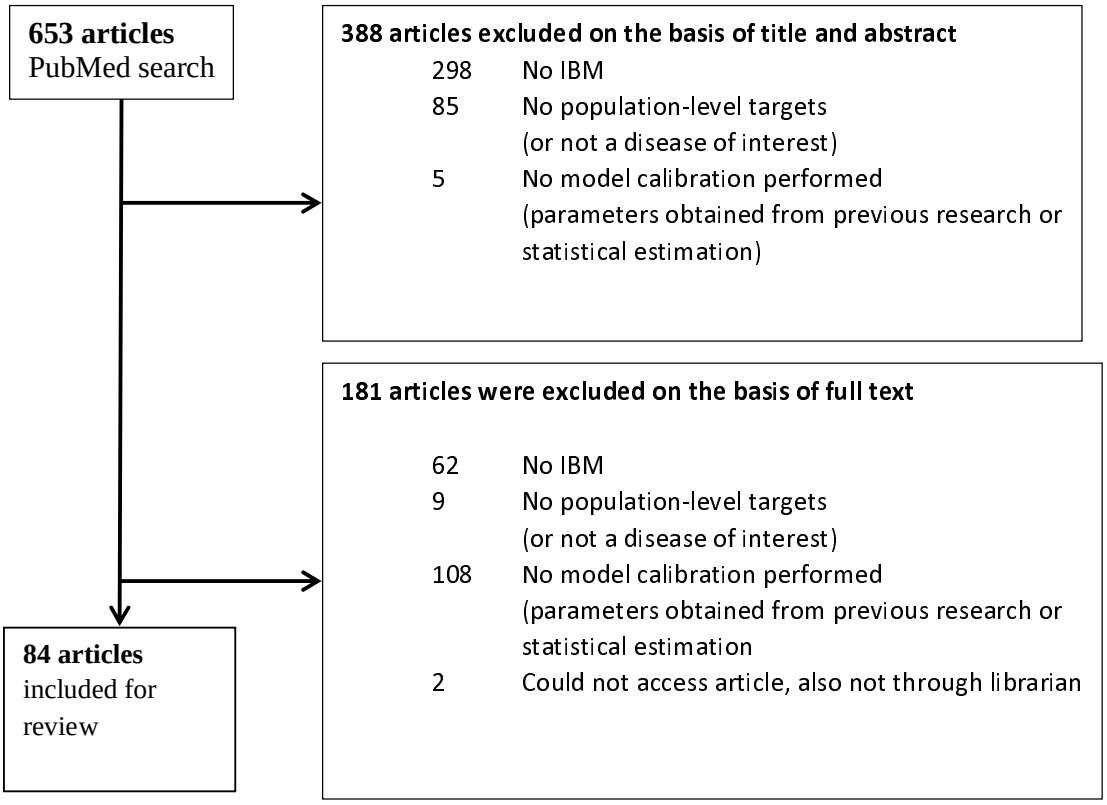

Fig 1. PRISMA flow diagram detailing the selection process of articles included in the review of individual-based models in epidemiology

Scope and objectives of included articles

The Table in S1 Table summarises the characteristics of the included articles. Fifty-eight (69%) of the included articles presented IBMs in HIV research, 16 (19%) concerned malaria, and another 10 (12%) concerned tuberculosis.

Most articles, namely 56 (67%), investigated the effect of an intervention, 17 articles looked at behavioural or biological explanations for the observed epidemic, and other goals (e.g. parameter estimation, model development) were used in 17. In total, six (7%) articles had two objectives. For most of these (five articles), one of the objectives was investigating the effect of an intervention (see Table A1).

Parameter-search strategies and measures of GOF

Of the included articles, 40 (48%) combined a quantitative measure of GOF with an algorithmic parameter-search strategy, which was an optimisation algorithm (14/40) or a sampling algorithm (26/40) (see Figure 2). For the remaining 44 articles, the parameter-search strategy could either not be identified (32/44) or was described as an informal, non-reproducible method (12/44).

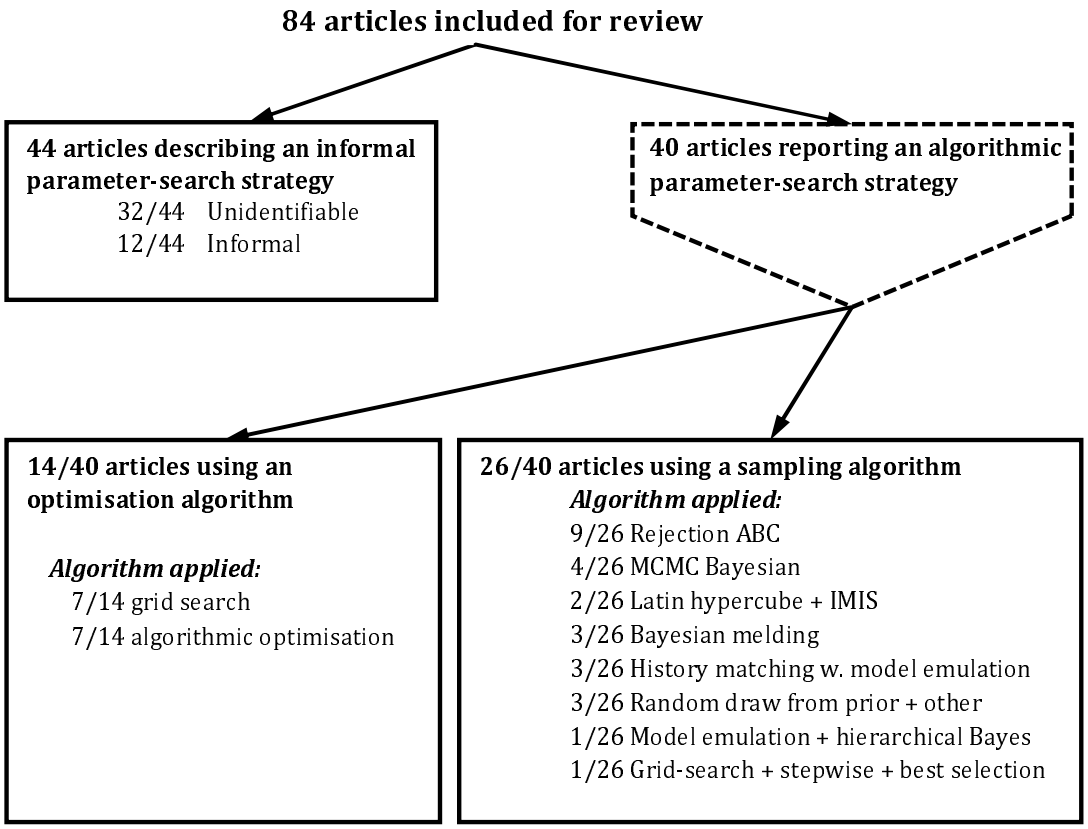

Fig 2. Reporting and application of parameter search strategies in epidemiological studies.

Detailed information on calibration methods for the 14 articles using optimisation algorithms is reported in Table 1. For the parameter-search strategy, most articles (9/14) used either a grid search (7), Latin square (1) or random draw from tolerable range (1), followed by the selection of the single best parameter combination. Several computer optimisation algorithms (i.e. Nelder-Mead, interior-point algorithm, coordinate descent with golden section search) were used in the remaining articles (5/14). Of these five articles, most (4/5) accepted a single best parameter combination without confidence intervals, while the remaining article

obtained confidence intervals around parameter estimates (see S2 Text.). For the GOF measure, the most common choice was a squared distance (5/14).

Table 1. Articles using optimisation algorithms for calibration.

Authors	Year	Pathogen	Parameter-search strategy	GOF
Bershteyn *et al*.	2013	HIV	Coordinate descent w. golden section search	Squared distance
Klein *et al*.	2015	HIV	Coordinate descent w. golden section search	Squared distance
Luo *et al*.	2018	HIV	Grid search	Absolute distance
Romero-Severson *et al*.	2013	HIV	Grid search	Kolmogorov-Smirnov
Marshall *et al*.	2018	HIV	Grid search	R-squared
Goedel *et al*.	2018	HIV	Grid search	R-squared and Manhattan distance of parameters
Brookmeyer *et al*.	2014	HIV	Grid search	Squared distance
Suen *et al*.	2014	TB	Grid search	Number of model outputs within the confidence intervals around the targets
Suen *et al*.	2015	TB	Grid search	Number of model outputs within the confidence intervals around the targets
Sauboin *et al*.	2015	Malaria	Interior point algorithm (hill-climbing alg.)	Squared distance
Shrestha *et al*.	2017	TB	Latin hypercube sampling	Likelihood
Knight *et al*.	2015	TB, HIV	Nelder-Mead	Squared distance
Kasaie *et al*.	2018	HIV	Random search mechanism	Absolute distance
Jewell *et al*.	2015	HIV	Sampling from tolerable range	Squared distance

Table 2 contains the details of the calibration methods in the 26 articles using sampling algorithms. Random sampling from the prior, followed by rejection ABC as a sampling algorithm, was used the most (9/26). Bayesian melding (3/26), history matching with model emulation (3/26), and MCMC (4/26) were also used. A variety of GOF measures were used, including chi-squares, relative distances, squared distances, and the approximate likelihood (see Table 2).

Table 2. Articles using sampling algorithms for calibration.

Authors	Year	Pathogen	Sampling algorithm	Sampling using GOF	GOF

Ciaranello *et al.*	2013	HIV	All possible combinations in grid	Acceptance	Absolute distance
Abuelezam *et al.*	2016	HIV	Bayesian melding	-	Squared distance
McCormick *et al.*	2014	HIV	Bayesian melding	-	Likelihood
McCormick *et al.*	2017	HIV	Bayesian melding	-	Likelihood
McCreesh *et al.*	2017	HIV	History matching w. model emulation	-	Implausibility measure
McCreesh *et al.*	2017	HIV	History matching w. model emulation	-	Implausibility measure
McCreesh *et al.*	2018	HIV	History matching w. model emulation	-	Implausibility measure
Huynh *et al.*	2015	TB	Latin hypercube	IMIS [b]	Likelihood
Chang *et al.*	2018	TB	Latin hypercube	IMIS [b]	Likelihood
Shcherbacheva *et al.*	2018	Malaria	MCMC [a]	-	Absolute distance
Penny *et al.*	2015	Malaria	MCMC [a]	-	Likelihood
Penny *et al.*	2015	Malaria	MCMC [a]	-	Likelihood
White *et al.*	2018	Malaria	MCMC [a]	-	Likelihood
Cameron *et al.*	2015	Malaria	Model emulation + hierarchical Bayes	-	Likelihood
Johnson *et al.*	2016	HIV	Random draw from prior	Best 500	Likelihood
Nakagawa *et al.*	2016	HIV	Random draw from prior	Rejection ABC	Relative distance
Nakagawa *et al.*	2017	HIV	Random draw from prior	Rejection ABC	Chi-square
Pizzitutti *et al.*	2018	Malaria	Random draw from prior	Rejection ABC	Squared distance
Cambiano *et al.*	2018	HIV	Random draw from prior	Rejection ABC	Relative distance
Hontelez *et al.*	2013	HIV	Random draw from prior	Rejection ABC	Squared distance
Phillips *et al.*	2013	HIV	Random draw from prior	Rejection ABC	Relative distance
Phillips *et al.*	2015	HIV	Random draw from prior	Rejection ABC	Relative distance
Shrestha *et al.*	2017	HIV	Random draw from prior	Rejection ABC	Absolute distance
Tuite *et al.*	2017	TB	Random draw from prior	Rejection ABC	Squared distance
Schalkwyk *et al.*	2018	HIV	Random draw from prior	SIR [c]	Likelihood
Pizzitutti *et al.*	2015	Malaria	Random draw from prior	Stepwise calibration	Absolute distance

[a] Markov chain Monte Carlo

[b] Incremental-mixture importance sampling (IMIS)

[c] Sampling importance resampling (SIR)

From the 44 articles with unidentifiable or informal parameter-search strategies, the majority (25 articles) are also unclear about the GOF used, while the rest either relied on visual inspection as a GOF (14 articles) or used a quantitative GOF (five articles).

Only 17 of the 84 included articles (20%) provided a rationale for their choice of parameter-search strategy. For example, McCreesh *et al*. [27] reported: "The model was fitted to the empirical data using history matching with model emulation, which allowed uncertainties in model inputs and outputs to be fully represented, and allowed realistic estimates of uncertainty in model results to be obtained" (see S3 Text. for more examples). Other examples indicate that an algorithmic calibration method failed to provide either a good fit or parameter estimates: "Ultimately, we chose to use visual inspection because the survival curves did not fit closely enough using the other two more quantitative approaches." [28] Or "[Calibration] was unable to resolve co-varying parameters. These parameters were adjusted by hand…" [29].

Ten out of the 84 articles included (12%) used a weighted calculation of GOF. Four articles weighted the GOF based on the amount of data behind the summary statistic fitted to, for example by weighting based on the inverse of the confidence interval around the data. In contrast, one article increased the weight for a data source for which fewer data was available. Other strategies included weighting based on a subjective assessment of the quality of the data, or weighting based on which data they wanted the model to fit best. One article down-weighted particular data to improve fit. Others stressed the importance of determining weights a priori since weights are chosen subjectively.

Acceptance criteria and stopping rules

None (0/14) of the articles applying optimisation algorithms mentioned the acceptance criteria or stopping rules. Acceptance criteria and stopping rules applied in studies using sampling algorithms can be summarised as running the model until obtaining an arbitrary number of accepted parameter combinations.

Target data and number of calibrated parameters

The number of target statistics varied from one to around 150. Target statistics were mostly obtained from registries or surveillance data; other sources included observational studies and, in a few cases, data from clinical trials. The number of calibrated parameters ranged from one to 96. The number of parameters for which the values were kept fixed, ranged from six to about 400. For several articles, the number of calibrated parameters appeared to exceed the number of target statistics. The size of the simulated population ranged from 100 to about 11 million and often reflected the size of the population of interest.

Computational aspects and the use of platforms

The software used to build the IBM was not reported in 34 (40%) of the articles. Seventeen articles (20%) used the low-level programming language C++, six (7%) used MATLAB, and another six (7%) used Python. A diverse range of other computing platforms were used in the remaining 23 (26%) articles. A high-performance computing facility was used in 16 (19%) articles.

Several simulation tools (i.e. CEPAC, EMOD, HIV-CDM, MicroCOSM, PATH, STDSIM and TITAN) were used in the articles modelling HIV. Similarly, a number of platforms (i.e. EMOD, OpenMalaria and Griffin IS) were used in the articles modelling malaria. In the articles modelling tuberculosis, the only tool reported was EMOD.

Model validation

Only 32 (38%) articles mentioned that a validation of the model had been performed.

Discussion

We conducted a systematic review of methods for calibrating IBMs in epidemiology. Of the included articles, two-thirds investigated the effects of interventions. When model results

are used to inform public health policy decisions, it is important to take into account uncertainty around the point estimates of the outcomes of interest, including the uncertainty introduced by calibrating unknown parameter values in the model [2,6-8,30,31]. For the majority of articles, either the parameter-search strategy could not be identified, or it was described as an informal, non-reproducible method. Less than half of the articles reviewed used algorithmic parameter-search strategies. Only one-third of the articles quantified the uncertainty around calibrated parameter values and could incorporate this uncertainty into their model projections. Articles that reported which formal calibration methods were used varied widely in their choice of parameter-search strategies and GOF measures. Only a few articles provided a rationale for their choice of parameter-search strategy. On occasion, the authors justified their choice of an informal method by indicating that algorithmic calibration methods did not converge to provide parameter estimates, or failed to provide a satisfactory fit to the targets. Only about 38% of articles reported on model validation.

Given the large number of articles for which we were unable to identify either parameter-search strategy or GOF, we encourage authors to use the standardised Calibration Reporting Checklist of Stout *et al.* [9]. While algorithmic parameter-search strategies favour reproducibility, unclear or incomplete reporting, and non-disclosure of software code, can still prevent reproducibility [32]. Manual adjustment of parameter values and visual inspection of GOF, on the other hand, does not favour reproducibility and involves subjective judgment, which may produce less than optimal calibration results and usually leads to the acceptance of a single parameter set (i.e. does not provide parameter uncertainty) [16]. Nevertheless, such informal methods may perform equally well compared to other methods in terms of GOF alone [33], may provide researchers with valuable insights into and familiarity with the model [34], and can be useful for simple didactic purposes [35-37].

There are several methodological challenges in the calibration of individual-based models, including the choice of calibration method – i.e. the combination of algorithmic

339 parameter-search strategy and GOF measure. The findings of the current review and previous
340 research suggest that there is no consensus on which calibration method to use [9,10,16,38,39].
341 Additionally, some of the articles reviewed here indicated that algorithmic calibration methods
342 had failed, leading the researchers to calibrate the model, either fully or partially, by hand. These
343 issues suggest that there is a need for research comparing the performance of calibration
344 methods to inform decisions about the choice of parameter-search strategy and GOF. Previous
345 research on calibration methods focused on the GOF [24], computation time and analyst time
346 [33]. Our review serves as a primer for methodological studies aimed at understanding the
347 relative performance of alternative calibration methods. Where applicable correct estimation of
348 the Bayesian posterior [40] should be a core aspect of performance, we further suggest
349 investigating several contextual variables, including the amount and nature of the empirical data
350 to calibrate against, the number and type of model parameters to be calibrated and insights to be
351 derived from the calibrated model. As evident from our review, these variables vary widely
352 across IBM studies in epidemiology.
353
354 Another methodological challenge in the calibration of IBMs is determining a priori
355 whether the target statistics provide sufficient information to calibrate the parameters [41].
356 Firstly, the target statistics are based on variable amounts of raw data. Secondly, a time series of
357 target statistics is often used, typically violating the assumption of independence that is often
358 implied by the calibration method. Thirdly, the complexity of the model may hamper the
359 specification of a prior parameter-distribution (including the specification of a correlation
360 between parameters) that is fully informed by prior knowledge of the data-generating processes
361 represented by the model. These problems preclude the use of standard statistical methods for
362 calculating the number of target statistics that is sufficient for parameter calibration. A related
363 problem is that target summary statistics are based on data from different sources, including
364 observational data that are potentially affected by treatment-confounder feedback [42].
365

366 The last methodological aspect of IBMs we would like to draw attention to is the size of
367 the simulated population [1,43]. Intuitively, one would recommend that the simulated
368 population size should be similar to the size of the population from which the samples were
369 drawn that gave rise to the target statistics. However, for many IBMs that aim to emulate hyper-
370 endemic epidemics of HIV, malaria or TB, such standards are beyond what is feasible with the
371 currently available computational infrastructure. Instead, researchers often adjust for the
372 inflated stochasticity in the modelled system by averaging outcomes of interest over multiple
373 simulation runs per parameter set [43]. How choices around modelled population size and
374 analysis of model output affect the validity of model inference deserves further attention in
375 future research.
376
377 Our results in the setting of HIV, TB and malaria IBMs indicate that the use of formal
378 calibration methods (49% of articles) is higher than in previous research on simulation models
379 in general – not IBMs specifically. Previously, only one-fifth to one-third of articles reporting on
380 epidemiological models used a quantitative GOF [9,44]. Our results concerning parameter
381 uncertainty are also optimistic compared to previous research by Stout *et al.* on calibration
382 methods in cancer models, which found that almost no articles quantified parameter
383 uncertainty, but instead accepted a single best-fitting parameter set as the result of the
384 calibration [9]. The same researchers reported that several different combinations of parameter-
385 search strategies and GOFs were used [9], outcomes which are similar to our findings. Stout *et al.*
386 report that articles rarely describe acceptance criteria and stopping rules. Stout *et al.* also report
387 that a standard description of the calibration process is lacking in almost all articles [9].
388 Similarly, previous research on IBMs of HIV transmission found that reporting was lacking in the
389 description of calibration methods [12]. All of this is in agreement with the results of the current
390 review. Concerning the goals of the included articles, our results agree with Punyacharoensin *et*
391 *al.* They found that the main goals of HIV-transmission models for the study of men who have
392 sex with men are: making projections for the epidemic, investigating how the incorporation of

various assumptions around the behavioural or biological characteristics affect these projections, and evaluating the impact of interventions [44].

To our knowledge, this is the first detailed review of methods used to calibrate IBMs of HIV, malaria and TB epidemics. An obvious limitation of our study is that we are unsure to what extent the results are generalisable to other infectious diseases. We encourage future research on other diseases to confirm or refute our current findings on the use of and reporting on methods in the calibration of IBMs in epidemiological research. On a related note, despite our best efforts, our PubMed search may have missed relevant articles by including the "NOT(molecular)" statement. However, we think there is no reason to assume that these would be a selective subset biasing the findings of this review. Another limitation comes from the fact that information on calibration is often scattered throughout the (often lengthy) IBM articles, including the appendixes. This might have caused us to miss information on calibration methods reported in the articles. We tried our best to avoid missing information by using two independent researchers instructed to search for information about calibration. A limitation of our review is that we treat articles that come from a given "research group" as independent observations. The calibration method used by a particular "research group" is usually the same across different articles. Treating each article as an independent observation may give a distorted view when we look at the percentage of articles in each of the categories. However, we aimed to give a full overview by providing information on individual articles in Tables 1 and 2. This information indicates that "modelling groups" tend to be inflexible in the way they calibrate their models to data, meaning that our independence assumption is potentially problematic.

In conclusion, it appears that calibrating individual-based models in epidemiological studies of HIV, malaria and TB transmission dynamics remains more of an art than a science. Besides limited reproducibility for a majority of the modelling studies in our review, our findings raise concerns over the correctness of model inference (e.g., estimated impact of past or future interventions) for models that are poorly calibrated. The quality of inference and reproducibility

in model-based epidemiology could benefit from the adoption of algorithmic parameter-search strategies and better-documented calibration and validation methods. We recommend the use of sampling algorithms to obtain valid estimates of parameter uncertainty and correlations between parameters. There is a need for research-based guidance on when to use which methods for calibrating IBMs to epidemiological data.

Acknowledgements

The authors gratefully acknowledge the help of all SACEMA students and researchers, specifically the fruitful conversations and helpful comments on the manuscript by Prof. Alex Welte, Mrs Cari van Schalkwyk, Dr Florian Marx, Prof. Juliet Pulliam and Dr Larisse Bolton. We would also like to acknowledge Mrs Marisa Honey and Mrs Susan Lotz from the Stellenbosch writing lab, who copy-edited the manuscript.

References

[1] Bobashev GV, Morris RJ. Uncertainty and inference in agent-based models. In: 2010 Second International Conference on Advances in System Simulation. IEEE; 2010. p. 67–71.

[2] Briggs AH, Weinstein MC, Fenwick EA, Karnon J, Sculpher MJ, Paltiel AD. Model parameter estimation and uncertainty analysis: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group–6. Medical decision making. 2012;32(5):722–732.

[3] Willem L, Verelst F, Bilcke J, Hens N, Beutels P. Lessons from a decade of individual-based models for infectious disease transmission: a systematic review (2006-2015). BMC infectious diseases. 2017;17(1):612.

[4] Hammond RA. Considerations and best practices in agent-based modeling to inform policy. In: Assessing the use of agent-based models for tobacco regulation. National Academies Press (US); 2015. .

[5] Johnson LF, Geffen N. A comparison of two mathematical modeling frameworks for evaluating sexually transmitted infection epidemiology. Sexually transmitted diseases. 2016;43(3):139–146.

[6] Kennedy MC, O'Hagan A. Bayesian calibration of computer models. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2001;63(3):425–464.

[7] Egger M, Johnson L, Althaus C, SchÃ¶ni A, Salanti G, Low N, et al. Developing WHO guidelines: Time to formally include evidence from mathematical modelling studies. F1000Research. 2017;6:1584.

[8] Menzies NA, Soeteman DI, Pandya A, Kim JJ. Bayesian methods for calibrating health policy models: a tutorial. Pharmacoeconomics. 2017;35(6):613–624.

[9] 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):533–545.

[10] Dahabreh IJ, Chan JA, Earley A, Moorthy D, Avendano EE, Trikalinos TA, et al. A Review of Validation and Calibration Methods for Health Care Modeling and Simulation. In: Modeling and Simulation in the Context of Health Technology Assessment: Review of Existing Guidance, Future Research Needs, and Validity Assessment [Internet]. Agency for Healthcare Research and Quality (US); 2017. .

[11] Caro JJ, Eddy DM, Kan H, Kaltz C, Patel B, Eldessouki R, et al. Questionnaire to assess relevance and credibility of modeling studies for informing health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. Value in health. 2014;17(2):174–182.

[12] Abuelezam NN, Rough K, Seage III GR. Individual-based simulation models of HIV transmission: reporting quality and recommendations. PloS one. 2013;8(9):e75624.

[13] Lintusaari J, Gutmann MU, Dutta R, Kaski S, Corander J. Fundamentals and recent developments in approximate Bayesian computation. Systematic biology. 2017;66(1):e66–e82.

[14] Hartig F, Calabrese JM, Reineking B, Wiegand T, Huth A. Statistical inference for stochastic simulation models–theory and application. Ecology letters. 2011;14(8):816–827.

[15] Leombruni R, Richiardi M. Why are economists sceptical about agent-based simulations? Physica A: Statistical Mechanics and its Applications. 2005;355(1):103–109.

[16] Vanni T, Karnon J, Madan J, White RG, Edmunds WJ, Foss AM, et al. Calibrating models in economic evaluation. Pharmacoeconomics. 2011;29(1):35–49.

[17] Nelder JA, Mead R. A simplex method for function minimization. The computer journal. 1965;7(4):308–313.

[18] Amaran S, Sahinidis NV, Sharda B, Bury SJ. Simulation optimization: a review of algorithms and applications. Annals of Operations Research. 2016;240(1):351–380.

[19] Joshi M, Seidel-Morgenstern A, Kremling A. Exploiting the bootstrap method for quantifying parameter confidence intervals in dynamical systems. Metabolic engineering. 2006;8(5):447–455.

[20] Stryhn H, Christensen J. Confidence intervals by the profile likelihood method, with applications in veterinary epidemiology. In: Proceedings of the 10th International Symposium on Veterinary Epidemiology and Economics, Vina del Mar; 2003. p. 208.

[21] McKinley TJ, Vernon I, Andrianakis I, McCreesh N, Oakley JE, Nsubuga RN, et al. Approximate Bayesian Computation and simulation-based inference for complex stochastic epidemic models. Statistical science. 2018;33(1):4–18.

[22] Rubin D. Using the SIR algorithm to simulate posterior distributions. Bayesian Stat. 1988;3:395–402.

[23] Poole D, Raftery AE. Inference for deterministic simulation models: the Bayesian melding approach. Journal of the American Statistical Association. 2000;95(452):1244–1255.

495 [24] Schunn CD, Wallach D, et al. Evaluating goodness-of-fit in comparison of models to data.
496 Psychologie der Kognition: Reden and vorträge anlässlich der emeritierung von Werner Tack.
497 2005;p. 115–154.

498 [25] Afzali HHA, Gray J, Karnon J. Model performance evaluation (validation and calibration)
499 in model-based studies of therapeutic interventions for cardiovascular diseases. Applied health
500 economics and health policy. 2013;11(2):85–93.

501 [26] Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic
502 reviews and meta-analyses: the PRISMA statement. Annals of internal medicine.
503 2009;151(4):264–269.

504 [27] McCreesh N, Andrianakis I, Nsubuga RN, Strong M, Vernon I, McKinley TJ, et al. Universal
505 test, treat, and keep: improving ART retention is key in cost-effective HIV control in Uganda.
506 BMC infectious diseases. 2017;17(1):322.

507 [28] Kessler J, Nucifora K, Li L, Uhler L, Braithwaite S. Impact and Cost-Effectiveness of
508 Hypothetical Strategies to Enhance Retention in Care within HIV Treatment Programs in East
509 Africa. Value in health : the journal of the International Society for Pharmacoeconomics and
510 Outcomes Research. 2015 dec;18(8):946–955. Available from: http://linkinghub.elsevier.com/-
511 retrieve/pii/S1098301515050731.

512 [29] Klein DJ, Eckhoff PA, Bershteyn A. Targeting HIV services to male migrant workers in
513 southern Africa would not reverse generalized HIV epidemics in their home communities: A
514 mathematical modeling analysis. International Health. 2015 mar;7(2):107–113.

515 [30] Gilbert JA, Meyers LA, Galvani AP, Townsend JP. Probabilistic uncertainty analysis of
516 epidemiological modeling to guide public health intervention policy. Epidemics. 2014;6:37–45.

517 [31] Caro JJ, Briggs AH, Siebert U, Kuntz KM. Modeling good research practicesâ€"overview: a
518 report of the ISPOR-SMDM Modeling Good Research Practices Task Force–1. Medical Decision
519 Making. 2012;32(5):667–677.

[32] Fehr J, Heiland J, Himpe C, Saak J. Best practices for replicability, reproducibility and reusability of computer-based experiments exemplified by model reduction software. arXiv preprint arXiv:160701191. 2016;.

[33] Taylor DC, Pawar V, Kruzikas D, Gilmore KE, Pandya A, Iskandar R, et al. Methods of model calibration. Pharmacoeconomics. 2010;28(11):995–1000.

[34] Gerberry DJ. An exact approach to calibrating infectious disease models to surveillance data: The case of HIV and HSV-2. Mathematical Biosciences & Engineering. 2018;15(1):153–179.

[35] Hodges JS. Six (or so) things you can do with a bad model. Operations Research. 1991;39(3):355–365.

[36] Kenyon CR, Delva W, Brotman RM. Differential sexual network connectivity offers a parsimonious explanation for population-level variations in the prevalence of bacterial vaginosis: a data-driven, model-supported hypothesis. BMC women's health. 2019;19(1):8.

[37] Delva W, Leventhal GE, Helleringer S. Connecting the dots: network data and models in HIV epidemiology. Aids. 2016;30(13):2009–2020.

[38] Karnon J, Vanni T. Calibrating models in economic evaluation. Pharmacoeconomics. 2011;29(1):51–62.

[39] Kopec JA, Finès P, Manuel DG, Buckeridge DL, Flanagan WM, Oderkirk J, et al. Validation of population-based disease simulation models: a review of concepts and methods. BMC public health. 2010;10(1):710.

[40] Talts S, Betancourt M, Simpson D, Vehtari A, Gelman A. Validating Bayesian inference algorithms with simulation-based calibration. arXiv preprint arXiv:180406788. 2018;.

[41] Srikrishnan V, Keller K. How much data are needed to calibrate and test agent-based models? arXiv preprint arXiv:181108524. 2018;.

[42] Murray EJ, Robins JM, Seage III GR, Lodi S, Hyle EP, Reddy KP, et al. Using observational data to calibrate simulation models. Medical Decision Making. 2018;38(2):212–224.

545 [43] Lee JS, Filatova T, Ligmann-Zielinska A, Hassani-Mahmooei B, Stonedahl F, Lorscheid I,
546 et al. The complexities of agent-based modeling output analysis. Journal of Artificial Societies
547 and Social Simulation. 2015;18(4):4.

548 [44] Punyacharoensin N, Edmunds WJ, De Angelis D, White RG. Mathematical models for the
549 study of HIV spread and control amongst men who have sex with men. European journal of
550 epidemiology. 2011;26(9):695.