
Article:
Mandrik, O. orcid.org/0000-0003-3755-3031, Thomas, C., Whyte, S. et al. (1 more author) (2022) Calibrating natural history of cancer models in the presence in the presence of data incompatibility; problems and solutions. PharmacoEconomics, 40 (4). pp. 359-366. ISSN 1170-7690

https://doi.org/10.1007/s40273-021-01125-3

CALIBRATING NATURAL HISTORY OF CANCER MODELS IN THE PRESENCE OF DATA INCOMPATIBILITY: PROBLEMS AND SOLUTIONS.

Type of paper: current opinion

Running title: data incompatibility in models

Olena Mandrik (0000-0003-3755-3031)[1], PhD; Chloe Thomas[1], PhD; Sophie Whyte[1], PhD; James Chilcott[1], Prof.

[1]Health Economic Modelling, School of Health and Related Research, University of Sheffield, Regent Court, Sheffield S1 4DA.

Corresponding author contact information:

Mandrik Olena

o.mandrik@sheffield.ac.uk

Kew words: 2,991

models development, microsimulation cancer model, models calibration, data compatibility

Acknowledgement: We are grateful to Prof. Mark Strong, Dr. Paul Tappenden, and Dr. Pete Dodd from the University of Sheffield for providing a critical feedback on the manuscript and to Dr. Nick Menzies for providing clarifications on his published work.

Abstract

Calibration of cancer natural history models is often challenged by a lack of representative calibration targets forcing modellers to rely on potentially incompatible datasets. Using a microsimulation colorectal cancer model as an example, the purposes of this paper are to: (1) highlight the reasons for uncertainty in calibration targets, (2) illustrate practical and generalizable approaches for dealing with incompatibility in calibration targets, and (3) discuss the importance of future research in the area of incorporating uncertainty in calibration.

Low quality of data and differences in populations, outcome definitions, and healthcare systems may result in incompatibility between the model and the data. Acknowledging reasons for data incompatibility allows assessment of the risk of incompatibility before calibrating the model. Only a few approaches are available to address data incompatibility, for instance addressing biases in calibration targets and their adjustment, relaxing the goodness-of-fit metric, and validation of the calibration targets to the data not used in the calibration. However, these approaches lack explicit comparison and validation and so more research is needed to describe the nature and causes of indirect uncertainty (i.e. uncertainty that cannot be expressed in absolute quantitative forms) and identify methods for managing this uncertainty in healthcare modelling.

Key points

- Extensive practical and theoretical recommendations discuss modelling parameter uncertainty; however, uncertainty related to the selection and compatibility of calibration targets is widely overlooked.
- A lack of representative calibration targets require models to rely on multiple datasets, which may not be compatible because of limited data quality, differences in definitions of outcomes, target populations, origin of the evidence, and/or healthcare settings.
- Data incompatibility in calibration of disease models may be addressed through adjustments of calibration targets, relaxing the goodness-of-fit metric used within the calibration algorithm, and validation of calibration targets to data not used in calibration.
- There is a critical need for standardising approaches for addressing incompatibility in calibration data; for instance through development of quantitative metrics to measure compatibility between a model and its calibration targets.

Declarations

Funding

Development of MiMiC-Bowel, its calibration and validation, was funded by the English National Screening Committee and Research England (Quality-related research funding to Support Evidence Based Policy Making distributed by the University of Sheffield). The views expressed are those of the authors and not necessarily those of the funding agencies.

Conflicts of interest/Competing interests

No conflict of interests to declare

Availability of data and material (data transparency)

The detailed description of the model used by the authors is reported in the referenced online reports. The model was populated using publicly available data.

Authors' contributions: OM wrote the draft manuscript, CT, SW and JC revised the draft manuscript. OM, CT, SW – co-developed the model used as a Worked Example.

1. Introduction

Cancer natural history models are used to assess the long-term impact and cost-effectiveness of preventive, screening, and treatment interventions. These models represent the natural history of cancer, including the onset of pre-cancerous conditions, cancer and cancer progression either by the definition of discrete health states or continuous growth models. In either case health state transition probabilities or cancer growth rates may not be directly observable as they occur in the asymptomatic population who are treated immediately if disease is detected. These unobserved parameters may be assessed through the computationally intensive statistical process called calibration. Calibration involves adjusting unknown model parameters to predict target statistics, called calibration targets. The calibration targets for cancer models frequently include cancer incidence and mortality, prevalence of pre-cancer states, and accuracy of screening tests [1-4].

The validity of model predictions relies heavily on the compatibility of data used in the model to the decision problem setting. The incompatibility of data to the model (i.e. to the decision problem setting) arises when the target data source is biased and non-representative of the population of interest. This means that the true difference between the modelled population and the population presented in calibration targets is not captured by sampling uncertainty because of an additional undefined uncertainty related to data compatibility. The compatibility of target data to the model is achieved using representative local data or data from other settings that are either generalizable by nature or appropriately transferred (i.e. adapted from other setting to apply in the modelled setting [5]). We will hereafter refer to this process as data adjustment.

Although multiple methodological reviews and empirical studies widely discuss parameter uncertainty (i.e. uncertainty in data used directly to populate the models), the uncertainty associated with the compatibility of data used to calibrate the models is rarely explored [6-10]. Requirements in the number of calibration targets are directly proportional to complexity of a model [1, 11]. Multiple calibration targets are required for successful parameterisation of highly complex models whose accuracy depends upon unobserved transitions between many health states [1]. Calibration of cancer natural history models faces a common challenge: a lack of representative calibration targets. The scarcity of reliable statistical data often forces modellers to rely on non-representative or potentially biased datasets and so on assumptions regarding compatibility of data sources. This problem of potential incompatibility between multiple calibration targets and models was recognised in previous research [12, 13]. Such biased incompatible datasets could include, for instance, data from different time periods, administrative units (e.g. regional instead of national data), or geographic settings (international data).

Incompatibility of target data to the model may be considered as an issue of indirect uncertainty (i.e. uncertainty that cannot be expressed in absolute quantitative forms [14]). There is some, albeit limited, evidence to suggest that knowledge of indirect uncertainty substantially affects the model-based decision-making [14]. This paper aimed to: (1) highlight the reasons for uncertainty in calibration targets, (2) illustrate a practical and generalizable approach for dealing with uncertainty in calibration targets, and (3) discuss the importance of future research in the area of incorporating uncertainty in calibration.

2. *Example model*

For the purposes of this paper we use an example model - the Microsimulation Model in Cancer of the Bowel (MiMiC-Bowel) that incorporates individual cancer risk to inform colorectal cancer (CRC) screening decisions in England [15]. The model was calibrated using the Metropolis-Hasting algorithm [16]. To ensure model identifiability, i.e. ability of the model to produce unique outputs for the calibrated parameters [11], and avoid under-specification, parameterisation of MiMiC-Bowel included 12 calibration targets: 5 based on English observational data (with longitudinal and completeness limitations), and 7 based on data from a German population study (with geographical, healthcare setting and data definition limitations). The calibration targets consisted of pre-screening CRC incidence total ($z_{i,}$) and by Duke's stage ($z_{i,a}$, $z_{i,b,}$ $z_{i,c,}$ $z_{i,d}$) in England and prevalence of low-risk (z_{la}) and high-risk (z_{ha}) adenomas and undiagnosed CRC total (z_u) and by stage from Germany ($z_{u,a}$, $z_{u,b,}$ $z_{u,c,}$ $z_{u,d}$) [17-20].

The calibration of the cancer natural history of the MiMiC-Bowel model assumed taking as inputs a set of parameters $x \in \{X \subseteq \mathbb{R}: \theta > 0 \wedge \theta < 1\}$ to produce an expected value of model outputs $y \in \{Y \subseteq \mathbb{R}\}$. This means that the modelling outputs (incidence of CRC total and by stages [$y_{i,}$ $y_{i,a}$, $y_{i,b,}$ $y_{i,c,}$ $y_{i,d}$], undiagnosed CRC total and by stages [$y_{u,}$ $y_{u,a}$, $y_{u,b,}$ $y_{u,c,}$ $y_{u,d}$], low-risk [y_{la}] and high-risk [y_{ha}] adenoma prevalence by age and sex) are functions of a set of calibrated parameters reflecting the speed of disease development and progression (x_v) and a set of fixed parameters (x_f), y=E[$f(x_v, x_f)$]. The predictions of the modelling outputs are compared to the calibration targets: z_i, $z_{i,a}$, $z_{i,b,}$ $z_{i,c,}$ $z_{i,d,}$ z_{ui}, $z_{u,a}$, $z_{u,b,}$ $z_{u,c,}$ $z_{u,d,}$ z_{la}, and z_{ha} by age and sex.

3. *Reasons for indirect uncertainty in calibration targets*

Incompatibility of calibration targets in cancer models could occur due to:

(1) Poor quality and incomplete data

Poor or incomplete data mean high risk of bias or uncertainty around the true epidemiological values. The reporting quality of epidemiological data varies broadly [21], with pre-screening data being especially underreported. For instance, in the example MiMiC-Bowel model, incidence of CRC by stage in England, reported in 1996-2004 in the UK Association of Cancer Registries (UKACR) dataset, had more than 40% of patients with unknown CRC stage at diagnosis (Table 1) [22].

(2) Compatibility of the outcomes by their definitions

Differences in definitions of cancer outcomes (either related to disease severity, staging, or a screening program performance [23-25]) are common and will inevitably lead to incompatibility between the model and the target cancer data. In the example model, differences in the reporting of pre-cancerous lesions (as either advanced adenoma or high-risk adenoma [HRA])[17, 19, 26] and in definitions of HRA and low-risk adenoma (LRA)[27, 28] impact compatibility between the data and MiMiC-Bowel (Table 1).

(3) Population differences

Differences in cancer statistical data, even among geographically comparable settings [29, 30], are driven by demographic differences and exposures to risk factors [29, 31], signifying data incompatibility. For example, data on prevalence of pre-cancer and CRC lesions in Germany would be a highly uncertain (and possibly incompatible) calibration target for the English CRC model. This is because population-level exposures to CRC risk factors, such as diet, physical activity and alcohol consumption [32] result in a substantial variability in pre-screening CRC incidence in both countries (Table 1) [18, 33]. Even when populations in calibration targets and the model are from the same setting, they may still be incompatible [34]. For instance, in MiMiC-Bowel the adjustment of definitions of pre-cancer states were based on the UK Flexible Sigmoidoscopy Screening randomized controlled trial (UKFSST), which was the only data source that reported detection of advanced and HRA [16]. However, the detection rates of CRC and HRA were higher in the trial compared with the subsequent Bowel Cancer Screening Programme (BCSP) in England and this discrepancy has been ascribed to the inclusion criteria and uptake characteristics [34, 35].

(4) Healthcare setting differences

Peculiarities of healthcare settings may limit compatibility of target datasets. For example, cancer prevalence, a frequently-used calibration target [36, 37], would require adjustments to country-specific data on cancer survival [38]. For some settings, cancer survival data may not be accessible. In MiMiC-Bowel, instead of CRC prevalence, prevalence of undiagnosed lesions was used as a calibration target (Table 1). The target's assessment was based on lesion detection rate from the German colonoscopy screening programme and published sensitivity rates [39]. However, the indirect uncertainty related to this calibration target remains because of variability in performance of screening and diagnostic tests across the jurisdictions and the studies [40, 41].

4. Approaches to address incompatibility of data

Selection of fully compatible calibration targets is always desirable but not always possible. Acknowledgment of data incompatibility allows identification of potential issues in using adjusted calibration targets. For simple models or for models where each unobserved model state is already informed by compatible target data [1, 11], supplementary incompatible data may be excluded from the calibration process. This though may rarely be the case in calibration of complex multi-state natural history disease models.

In some models, the uncertainty in calibration targets may be addressed by generating parameter distributions within the calibration process through bias adjustment [46-48]. If multiple biased studies are available, the bias-adjusted means and standard errors of target data can be combined in a bias-adjusted meta-analysis [49]. An explicit modelling of the biases (i.e. capturing both sampling uncertainty and uncertainty related to data incompatibility) would probably be the most robust method dealing with the data incompatibility. It may be challenging to apply though when biases around a calibration target are not well-understood [37], especially when there are uncertain biases around multiple calibration targets, either risking sampling inefficiencies if heavy tailed prior distributions are used [50] or more uncertain and so less informative for healthcare decisions.

Different data adjustment methods may either improve or resolve incompatibility of calibration targets to the model. If uncertainty related to data incompatibility can be measured, adjustment of calibration targets may

fully address the differences in definitions, populations, samples, or geographic and healthcare settings. However, an absence of metrics to quantify data compatibility [1, 51] makes it difficult to transform qualitative knowledge into informed quantitative data adjustment in unbiased way. Uncertainty around a calibration target (for either statistical or ad-hoc adjustment) may be informed by supplementary evidence, such as differences in disease incidence or detection rates across the settings, or through elicitation for data compatibility biases [49].

Lack of validity for data adjustment processes means that even after adjustment is applied to incompatible data, the risk of bias in some of these targets may still exist. If it is a case, such calibration targets may be weakened under ad-hoc or quasi-Bayesian framework [37]. Under accept-reject sampling schemes, different tolerances can be defined for various calibration targets [48, 52]. Within a goodness-of-fit metric, lower weights may be assigned to calibration targets with higher uncertainty. The convergence to individual datasets with the selected weights may be tested in a sensitivity analysis. While arbitrary weighting of target data in calibration is common [13, 51, 53], more objectivity can be given by involving an expert elicitation. This process involves asking the clinical experts and the decision-makers about their confidence in adjusted data, by inquiring about the probability for datasets to represent the modelled population. The elicitation approach may involve individual or group experts' consultations (sometimes also relying on elicitation scales), and ideally is based on a structural protocol [54].

Independently of any former steps undertaken, compatibility of the calibration targets may be assessed before undertaking the calibration process. One- and multi-way way sensitivity analyses can assess compatibility of different calibration targets within the fixed model structure and can be used to signal incompatibility issues. Another way would be to perform an external validation of calibration targets (e.g. feasibility of the predicted accuracy of screening when natural history disease parameters are calibrated) or a cross validation to outputs predicted by other models (e.g. sojourn time).

5. *Incompatibility of data in MiMiC-Bowel*

To illustrate possible data adjustment in regard to different compatibility problems, we present some examples from dealing with data incompatibility in MiMiC-Bowel (see the report for full details [16]). The complexity of a model with multiple unobserved transitions required multiple calibration targets for parametrisation, and so incompatible calibration targets could not be excluded from the calibration.

5.1. Data adjustment to increase data compatibility

In MiMiC-Bowel, the calibration target data for pre-screening CRC incidence by stage were adjusted (using an additional data set) to address the poor data completeness issue. The unstaged cases were distributed between stages C and D so that predicted patient survival, where total survival S_o is a function of proportion of population in each disease stage at diagnosis and a survival at each stage ($S_{a,b,c,d}$), $S_o = f(P_{a,b,c,d}; S_{a,b,c,d})$, where stage-specific survival ($S_{a,b,c,d}$) is a function of age, sex, and time since diagnosis: $S_{a,b,c,d} = f(age; sex, t_{diag})$ matched that in observed data [15].

Calibration of MiMiC-Bowel was hampered by the different outcome definitions used in potential target datasets ; thus, adjustments of calibration targets to reflect definitions used in the model were applied when necessary. Stage distribution of undiagnosed CRC was based on a German multi-centre cohort study in 2003-2010 [20]. In this case, no data adjustments were applied after comparison of the definitions of stages I-IV in Brenner *et. al*. (2016) and Dukes' classification due to their similarity and an absence of a direct guide for case conversion [20, 55]. However, the definitions of pre-cancer lesions were different in the model and the calibration target [17]. While MiMiC-Bowel assessed cancer progression through low- and high-risk adenomas, the German colonoscopy screening programme reported detection rates for advanced and non-advanced adenomas [17]. The adjustment relied on information from the UKFSST, the only source found reporting detection rates for both advanced adenomas and HRA [17, 26]. To adjust the target, prevalence of all adenomas was calculated as: $Prev_a = \frac{DR(aa)}{Sens\ (aa)} + \frac{DR(na)}{Sens\ (na)}$, where DR - detection rates of advanced and non-advanced adenomas reported by Brenner *et. al*. [17], and Sens - sensitivity of colonoscopy to advanced and non-advanced adenomas reported in the literature [44, 56]. The prevalence of HRA was calculated by multiplying calculated prevalence of advanced adenoma by the proportion of advanced adenomas that are high-risk among males $[P\left(\frac{aa}{HRA}; men\right)]$ and females $[P\left(\frac{aa}{HRA}; women\right)]$ in UKFSST, $Prev_{HRA} = \frac{DR(aa,men)}{Sens\ (aa)} \times P\left(\frac{aa}{HRA}; men\right) + \frac{DR(na,women)}{Sens\ (na)} \times P\left(\frac{aa}{HRA}; women\right)$. The proportion of advanced adenomas that are high-risk in UKFSST was 0.81 for males and 0.66 for females [16]. Prevalence of LRA then was calculated as a difference between total adenoma prevalence and prevalence of HRA, $Prev_{LRA} = Prev_a - Prev_{HRA}$.

Adjustment of calibration targets because of incompatibility in geographic and healthcare settings was conducted in MiMiC-Bowel for undiagnosed CRC. Observing that pre-screening CRC incidence in England was 10% lower than in Germany among population of screening ages [16], we adjusted the prevalence of undiagnosed cancer in Germany with a 10% reduction assuming a correlation between these two outcomes.

5.2. Relaxing goodness-of-fit metric

All except one adjusted calibration targets ($z_{i,a}$, $z_{i,b}$, $z_{i,c}$, $z_{i,d}$, z_{ui}, $z_{u,a}$, $z_{u,b}$, $z_{u,c}$, $z_{u,d}$, z_{la}, and z_{ha}) were assumed to have additional uncaptured uncertainty related to data incompatibility. Thus, in the calibrated MiMiC-Bowel model we accepted lower precision in adjusted targets by giving them lower weights in the calibration. The total SSE in the calibration was equal to the sum of weighted SSE for each calibration target i, $Total\ SSE = \left(\sum_{i=1}^{\#\ data\ targets} weight_i \times SSE_i\right)$. The selected approach resulted in the calibration algorithm prioritising fitting to total CRC incidence [z_i] by age and sex in the population over the other calibration targets [16].

5.3. *Validation of calibrated model parameters against data not used in the calibration*

In the example MiMiC-Bowel model we assessed compatibility of calibration targets by validating them to sensitivity of faecal immunochemical test with positivity threshold of 20 µg haemoglobin/g feces (FIT20) and flexible sigmoidoscopy screening (FS) screening tests in England to HRA and CRC. The validation assumed that the upper bound for sensitivity of FS in population screening can reach 72% for CRC and 69% for HRA. This threshold was retrieved by firstly calculating the maximum possible sensitivity of FS that could be obtained

in a trial setting ($FS_{trial,\ sens}$), using test sensitivity to distal (S_{dl}) and proximal lesions (S_{pl}) reported in the literature [44, 56] and proportions of CRC and advanced adenomas that are distal (P_{dl}) and proximal (P_{pl}) from UKFSST data [34, 57]: $FS_{UKFSST,\ sens} = S_{pl} \times P_{pl} + S_{dl} \times P_{dl}$. The predicted sensitivity of FS in the UKFSST to CRC and HRA was adjusted to reflect the lower performance of the national screening programme [16] ($FS_{BCSP,\ sens}$), $FS_{\text{BCSP, sens}} = \frac{DR(BCSP)}{DR\ (UKFSST)} \times FS_{\text{UKFSST,}sens.}$

For FIT20, validation assumed an upper bound of sensitivity to CRC of 52% based on trial data [58]. Sensitivity of tests was calculated by dividing detection rates in BCSP and FIT pilot by prevalence of lesions estimated with the target calibration datasets. With adjusted calibration targets, FS sensitivity was 60% for CRC and 59% for HRA for 55-year olds. The predicted FIT20 sensitivity in BCSP was 48% for CRC and 32% for HRA for 60-year olds. The predicted FIT20 sensitivity with non-adjusted calibration targets (without adjustments to cross-country differences between England and Germany, as described above) was above the calculated threshold.

6. Limitations and future research considerations

Adjusting calibration targets enables increased data compatibility within the model. In the absence of explicit comparisons of different methods addressing incompatibility in calibration targets, ad-hoc adjustments have been applied. The uncertainty in data adjustments increases when multiple adjustment steps are needed (e.g. when multiple incompatibility reasons are identified), accumulating uncertainty related to each step. Thus, while multiple reasons for incompatibility were identified for some of the calibration targets for MiMiC-Bowel, the adjustment steps were limited to those subjectively considered to be the most influential, while increasing the uncertainty around the calibration targets by assigning less weight to uncertain datasets in the calibration process. While arbitrary weighting of target data in calibration is common, more empirical research is needed to explore the impact of assigning weights on uncertainty and bias.

Indirect uncertainty in modelling remains high ahead of development of a standardized process for calibration target selection and adjustment. Significant controversy regarding compatibility of calibration targets highlights a critical need to develop a quantitative metric to identify and measure incompatibility and to adjust calibration data, as well as to compare available methods aiming to address data incompatibility. More research is needed to identify markers for data incompatibility that could be used systematically in models' calibration. Similarly, there is a need to quantify indirect uncertainty in healthcare modelling and to understand the impact of indirect uncertainty on healthcare decision-making. While the available recommendations on models calibration focus on technical issues [36, 37, 51, 59], we call for guidance on data compatibility in healthcare models calibration.

7. *References*

1. 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. Epub 2010/12/15. doi: 10.2165/11584600-000000000-00000. PubMed PMID: 21142277.
2. Platt D. A comparison of economic agent-based model calibration methods. Journal of Economic Dynamics and Control. 2020;113:103859. doi: https://doi.org/10.1016/j.jedc.2020.103859.
3. Whyte S, Walsh C, Chilcott J. Bayesian calibration of a natural history model with application to a population model for colorectal cancer. Med Decis Making. 2011;31(4):625-41. Epub 2010/12/04. doi: 10.1177/0272989x10384738. PubMed PMID: 21127321.
4. 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. Epub 2009/08/12. doi: 10.2165/11314830-000000000-00000. PubMed PMID: 19663525; PubMed Central PMCID: PMCPMC2787446.
5. Drummond M, Barbieri M, Cook J, Glick HA, Lis J, Malik F, et al. Transferability of economic evaluations across jurisdictions: ISPOR Good Research Practices Task Force report. Value Health. 2009;12(4):409-18. Epub 2009/11/11. doi: 10.1111/j.1524-4733.2008.00489.x. PubMed PMID: 19900249.
6. Corro Ramos I, Hoogendoorn M, Rutten-van Mölken MPMH. How to Address Uncertainty in Health Economic Discrete-Event Simulation Models: An Illustration for Chronic Obstructive Pulmonary Disease. Medical Decision Making. 2020;40(5):619-32. doi: 10.1177/0272989x20932145. PubMed PMID: 32608322.
7. D'Agostino McGowan L, Grantz KH, Murray E. Quantifying Uncertainty in Mechanistic Models of Infectious Disease. American Journal of Epidemiology. 2021;190(7):1377-85. doi: 10.1093/aje/kwab013.
8. Bilcke J, Beutels P, Brisson M, Jit M. Accounting for methodological, structural, and parameter uncertainty in decision-analytic models: a practical guide. Med Decis Making. 2011;31(4):675-92. Epub 2011/06/10. doi: 10.1177/0272989x11409240. PubMed PMID: 21653805.
9. Degeling K, IJzerman MJ, Koopman M, Koffijberg H. Accounting for parameter uncertainty in the definition of parametric distributions used to describe individual patient variation in health economic models. BMC Med Res Methodol. 2017;17(1):170. Epub 2017/12/17. doi: 10.1186/s12874-017-0437-y. PubMed PMID: 29246192; PubMed Central PMCID: PMCPMC5732462.
10. Briggs AH, Weinstein MC, Fenwick EA, Karnon J, Sculpher MJ, Paltiel AD. 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. Epub 2012/09/25. doi: 10.1016/j.jval.2012.04.014. PubMed PMID: 22999133.
11. Alarid-Escudero F, MacLehose RF, Peralta Y, Kuntz KM, Enns EA. Nonidentifiability in Model Calibration and Implications for Medical Decision Making. Med Decis Making. 2018;38(7):810-21. Epub 2018/09/25. doi: 10.1177/0272989x18792283. PubMed PMID: 30248276; PubMed Central PMCID: PMCPMC6156799.
12. Rutter CM, Ozik J, DeYoreo M, Collier N. Microsimulation model calibration using incremental mixture approximate bayesian computation. Ann Appl Stat. 2019;13(4):2189-212. Epub 2019/12/01. doi: 10.1214/19-aoas1279. PubMed PMID: 34691351; PubMed Central PMCID: PMCPMC8534811.
13. Kong CY, McMahon PM, Gazelle GS. Calibration of disease simulation model using an engineering approach. Value Health. 2009;12(4):521-9. Epub 2009/11/11. doi: 10.1111/j.1524-4733.2008.00484.x. PubMed PMID: 19900254; PubMed Central PMCID: PMCPMC2889011.
14. Padilla LMK, Powell M, Kay M, Hullman J. Uncertain About Uncertainty: How Qualitative Expressions of Forecaster Confidence Impact Decision-Making With Uncertainty Visualizations. Front Psychol. 2021;11:579267-. doi: 10.3389/fpsyg.2020.579267. PubMed PMID: 33564298.
15. Thomas C, Mandrik, O. and Whyte, S. Development of the Microsimulation Model in Cancer of the Bowel (MiMiC-Bowel), an Individual Patient Simulation Model for Investigation of the Cost-

effectiveness of Personalised Screening and Surveillance Strategies. 2020 1 April 2020. Report No. Available from: https://eprints.whiterose.ac.uk/162743/
16. Mandrik O. TC, Strong M. , Whyte S. Calibration and Validation of the Microsimulation Model in Cancer of the Bowel (MiMiC-Bowel), an Individual Patient Simulation Model for Investigation of the Cost-effectiveness of Personalised Screening and Surveillance Strategies. Sheffield: School of Health and Related Research, University of Sheffield, 2021. Available from: https://eprints.whiterose.ac.uk/171343/
17. Brenner H, Altenhofen L, Hoffmeister M. Sex, age, and birth cohort effects in colorectal neoplasms: a cohort analysis. Ann Intern Med. 2010;152(11):697-703. Epub 2010/06/02. doi: 10.7326/0003-4819-152-11-201006010-00002. PubMed PMID: 20513827.
18. Brenner H, Altenhofen L, Katalinic A, Lansdorp-Vogelaar I, Hoffmeister M. Sojourn time of preclinical colorectal cancer by sex and age: estimates from the German national screening colonoscopy database. Am J Epidemiol. 2011;174(10):1140-6. Epub 2011/10/11. doi: 10.1093/aje/kwr188. PubMed PMID: 21984657.
19. Brenner H, Altenhofen L, Stock C, Hoffmeister M. Incidence of colorectal adenomas: birth cohort analysis among 4.3 million participants of screening colonoscopy. Cancer Epidemiol Biomarkers Prev. 2014;23(9):1920-7. Epub 2014/07/12. doi: 10.1158/1055-9965.Epi-14-0367. PubMed PMID: 25012996.
20. Brenner H, Jansen L, Ulrich A, Chang-Claude J, Hoffmeister M. Survival of patients with symptom- and screening-detected colorectal cancer. Oncotarget. 2016;7(28):44695-704. Epub 2016/05/24. doi: 10.18632/oncotarget.9412. PubMed PMID: 27213584; PubMed Central PMCID: PMCPMC5190129.
21. Altobelli E, D'Aloisio F, Angeletti PM. Colorectal cancer screening in countries of European Council outside of the EU-28. World J Gastroenterol. 2016;22(20):4946-57. Epub 2016/05/31. doi: 10.3748/wjg.v22.i20.4946. PubMed PMID: 27239121; PubMed Central PMCID: PMCPMC4873887.
22. Incidence numbers of Colorectal Cancer for patients diagnosed between 1996 and 2004 in England , by stage. In: Registries UAoC, editor. 2009.
23. Kim SH, Shin DW, Kim SY, Yang HK, Nam E, Jho HJ, et al. Terminal Versus Advanced Cancer: Do the General Population and Health Care Professionals Share a Common Language? Cancer Res Treat. 2016;48(2):759-67. Epub 2015/08/10. doi: 10.4143/crt.2015.124. PubMed PMID: 26323640.
24. Mandrik O, Tolma E, Zielonke N, Meheus F, Ordóñez-Reyes C, Severens JL, et al. Systematic reviews as a “lens of evidence”: Determinants of participation in breast cancer screening. Journal of Medical Screening. 2020:0969141320930743. doi: 10.1177/0969141320930743.
25. Walters S, Maringe C, Butler J, Brierley JD, Rachet B, Coleman MP. Comparability of stage data in cancer registries in six countries: lessons from the International Cancer Benchmarking Partnership. Int J Cancer. 2013;132(3):676-85. Epub 2012/05/25. doi: 10.1002/ijc.27651. PubMed PMID: 22623157.
26. Atkin W, Wooldrage K, Parkin DM, Kralj-Hans I, MacRae E, Shah U, et al. Long term effects of once-only flexible sigmoidoscopy screening after 17 years of follow-up: the UK Flexible Sigmoidoscopy Screening randomised controlled trial. Lancet. 2017;389(10076):1299-311. Epub 2017/02/27. doi: 10.1016/s0140-6736(17)30396-3. PubMed PMID: 28236467; PubMed Central PMCID: PMCPMC6168937.
27. Winawer SJ, Zauber AG, Fletcher RH, Stillman JS, O'Brien MJ, Levin B, et al. Guidelines for colonoscopy surveillance after polypectomy: a consensus update by the US Multi-Society Task Force on Colorectal Cancer and the American Cancer Society. Gastroenterology. 2006;130(6):1872-85. Epub 2006/05/16. doi: 10.1053/j.gastro.2006.03.012. PubMed PMID: 16697750.
28. East JE, Atkin WS, Bateman AC, Clark SK, Dolwani S, Ket SN, et al. British Society of Gastroenterology position statement on serrated polyps in the colon and rectum. Gut. 2017;66(7):1181-96. Epub 2017/04/30. doi: 10.1136/gutjnl-2017-314005. PubMed PMID: 28450390; PubMed Central PMCID: PMCPMC5530473.
29. Torre LA, Siegel RL, Ward EM, Jemal A. Global Cancer Incidence and Mortality Rates and Trends--An Update. Cancer Epidemiol Biomarkers Prev. 2016;25(1):16-27. Epub 2015/12/17. doi: 10.1158/1055-9965.Epi-15-0578. PubMed PMID: 26667886.

30. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394-424. Epub 2018/09/13. doi: 10.3322/caac.21492. PubMed PMID: 30207593.
31. Wild CP, Espina C, Bauld L, Bonanni B, Brenner H, Brown K, et al. Cancer Prevention Europe. Mol Oncol. 2019;13(3):528-34. Epub 2019/01/23. doi: 10.1002/1878-0261.12455. PubMed PMID: 30667152; PubMed Central PMCID: PMCPMC6396376.
32. Keum N, Giovannucci E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. Nat Rev Gastroenterol Hepatol. 2019;16(12):713-32. Epub 2019/08/29. doi: 10.1038/s41575-019-0189-8. PubMed PMID: 31455888.
33. Cancer Registration Statistics, England [Internet]. 2005. Available from: https://webarchive.nationalarchives.gov.uk/20160307140012/https://cy.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/datasets/cancerregistrationstatisticscancerregistrationstatisticsengland.
34. Brown JP, Wooldrage K, Kralj-Hans I, Wright S, Cross AJ, Atkin WS. Effect of once-only flexible sigmoidoscopy screening on the outcomes of subsequent faecal occult blood test screening. J Med Screen. 2019;26(1):11-8. Epub 2018/10/05. doi: 10.1177/0969141318785654. PubMed PMID: 30282520; PubMed Central PMCID: PMCPMC6376653.
35. Siau K, Yew AC, Ishaq S, Jewes S, Shetty S, Brookes M, et al. Colonoscopy conversion after flexible sigmoidoscopy screening: results from the UK Bowel Scope Screening Programme. Colorectal Dis. 2018;20(6):502-8. Epub 2017/12/06. doi: 10.1111/codi.13982. PubMed PMID: 29205835.
36. 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. Epub 2013/07/28. doi: 10.1177/0272989x13493143. PubMed PMID: 23886677; PubMed Central PMCID: PMCPMC4847637.
37. Menzies NA, Soeteman DI, Pandya A, Kim JJ. Bayesian Methods for Calibrating Health Policy Models: A Tutorial. Pharmacoeconomics. 2017;35(6):613-24. Epub 2017/03/02. doi: 10.1007/s40273-017-0494-4. PubMed PMID: 28247184; PubMed Central PMCID: PMCPMC5448142.
38. Bray F, Ren J-S, Masuyer E, Ferlay J. Global estimates of cancer prevalence for 27 sites in the adult population in 2008. International Journal of Cancer. 2013;132(5):1133-45. doi: https://doi.org/10.1002/ijc.27711.
39. Bressler B, Paszat LF, Chen Z, Rothwell DM, Vinden C, Rabeneck L. Rates of new or missed colorectal cancers after colonoscopy and their risk factors: a population-based analysis. Gastroenterology. 2007;132(1):96-102. Epub 2007/01/24. doi: 10.1053/j.gastro.2006.10.027. PubMed PMID: 17241863.
40. Gies A, Cuk K, Schrotz-King P, Brenner H. Direct Comparison of Diagnostic Performance of 9 Quantitative Fecal Immunochemical Tests for Colorectal Cancer Screening. Gastroenterology. 2018;154(1):93-104. Epub 2017/09/30. doi: 10.1053/j.gastro.2017.09.018. PubMed PMID: 28958859.
41. Quyn AJ, Fraser CG, Stanners G, Carey FA, Rees CJ, Moores B, et al. Scottish Bowel Screening Programme colonoscopy quality - scope for improvement? Colorectal Dis. 2018;20(9):O277-o83. Epub 2018/06/05. doi: 10.1111/codi.14281. PubMed PMID: 29863812.
42. Bretthauer M, Kaminski MF, Loberg M, Zauber AG, Regula J, Kuipers EJ, et al. Population-Based Colonoscopy Screening for Colorectal Cancer: A Randomized Clinical Trial. JAMA Intern Med. 2016;176(7):894-902. Epub 2016/05/24. doi: 10.1001/jamainternmed.2016.0960. PubMed PMID: 27214731; PubMed Central PMCID: PMCPMC5333856.
43. van Rijn AF, Dekker E, Kleibeuker JH. [Screening the population for colorectal cancer: the background to a number of pilot studies in the Netherlands]. Ned Tijdschr Geneeskd. 2006;150(50):2739-44. Epub 2007/01/18. PubMed PMID: 17225784.
44. Martin-Lopez JE, Beltran-Calvo C, Rodriguez-Lopez R, Molina-Lopez T. Comparison of the accuracy of CT colonography and colonoscopy in the diagnosis of colorectal cancer. Colorectal Dis. 2014;16(3):O82-9. Epub 2013/12/05. doi: 10.1111/codi.12506. PubMed PMID: 24299052.
45. Census. Office for National Statistics. [Internet]. Office for National Statistics. . 2005. Available from: https://www.ons.gov.uk/search?q=2005+census.

46. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004;15(5):615-25. Epub 2004/08/17. doi: 10.1097/01.ede.0000135174.63482.43. PubMed PMID: 15308962.
47. Sauboin CJ, Van Bellinghen L-A, Van De Velde N, Van Vlaenderen I. Potential public health impact of RTS,S malaria candidate vaccine in sub-Saharan Africa: a modelling study. Malar J. 2015;14:524-. doi: 10.1186/s12936-015-1046-z. PubMed PMID: 26702637.
48. Ward ZJ, Yeh JM, Bhakta N, Frazier AL, Girardi F, Atun R. Global childhood cancer survival estimates and priority-setting: a simulation-based analysis. The Lancet Oncology. 2019;20(7):972-83. doi: https://doi.org/10.1016/S1470-2045(19)30273-6.
49. Turner RM, Lloyd-Jones M, Anumba DOC, Smith GCS, Spiegelhalter DJ, Squires H, et al. Routine antenatal anti-D prophylaxis in women who are Rh(D) negative: meta-analyses adjusted for differences in study design and quality. PLoS One. 2012;7(2):e30711-e. Epub 2012/02/03. doi: 10.1371/journal.pone.0030711. PubMed PMID: 22319580.
50. König C, Spoden C, Frey A. An Optimized Bayesian Hierarchical Two-Parameter Logistic Model for Small-Sample Item Calibration. Appl Psychol Meas. 2020;44(4):311-26. Epub 2019/12/21. doi: 10.1177/0146621619893786. PubMed PMID: 32536732.
51. Karnon J, Vanni T. Calibrating models in economic evaluation: a comparison of alternative measures of goodness of fit, parameter search strategies and convergence criteria. Pharmacoeconomics. 2011;29(1):51-62. Epub 2010/12/15. doi: 10.2165/11584610-000000000-00000. PubMed PMID: 21142278.
52. Kypraios T, Neal P, Prangle D. A tutorial introduction to Bayesian inference for stochastic epidemic models using Approximate Bayesian Computation. Mathematical Biosciences. 2017;287:42-53. doi: https://doi.org/10.1016/j.mbs.2016.07.001.
53. Taylor DC, Pawar V, Kruzikas D, Gilmore KE, Pandya A, Iskandar R, et al. Methods of model calibration: observations from a mathematical model of cervical cancer. Pharmacoeconomics. 2010;28(11):995-1000. Epub 2010/10/13. doi: 10.2165/11538660-000000000-00000. PubMed PMID: 20936883.
54. Hemming V, Burgman MA, Hanea AM, McBride MF, Wintle BC. A practical guide to structured expert elicitation using the IDEA protocol. Methods in Ecology and Evolution. 2018;9(1):169-80. doi: https://doi.org/10.1111/2041-210X.12857.
55. Rudy DR, Zdon MJ. Update on colorectal cancer. Am Fam Physician. 2000;61(6):1759-70, 73-4. Epub 2000/04/06. PubMed PMID: 10750881.
56. Castro I, Estevez P, Cubiella J, Hernandez V, Gonzalez-Mao C, Rivera C, et al. Diagnostic performance of fecal immunochemical test and sigmoidoscopy for advanced right-sided colorectal neoplasms. Dig Dis Sci. 2015;60(5):1424-32. Epub 2014/11/20. doi: 10.1007/s10620-014-3434-6. PubMed PMID: 25407805.
57. Brenner H, Niedermaier T, Chen H. Strong subsite-specific variation in detecting advanced adenomas by fecal immunochemical testing for hemoglobin. Int J Cancer. 2017;140(9):2015-22. Epub 2017/02/06. doi: 10.1002/ijc.30629. PubMed PMID: 28152558.
58. Niedermaier T, Tikk K, Gies A, Bieck S, Brenner H. Sensitivity of Fecal Immunochemical Test for Colorectal Cancer Detection Differs According to Stage and Location. Clin Gastroenterol Hepatol. 2020. Epub 2020/01/29. doi: 10.1016/j.cgh.2020.01.025. PubMed PMID: 31988043.
59. Afzali HH, Karnon J. Exploring structural uncertainty in model-based economic evaluations. Pharmacoeconomics. 2015;33(5):435-43. Epub 2015/01/21. doi: 10.1007/s40273-015-0256-0. PubMed PMID: 25601288.

Table 1 Reasons for data and model incompatibility and approaches applied to address the issue

Calibration target	Data sources considered	Potential causes of incompatibility	Data target(s) used and reason	Approach(es) taken to address possible data incompatibility
Low risk adenomas (LRA) and High risk adenomas (HRA) prevalence by age and sex	RCT in Norway, Sweden, Netherlands, Poland[42]	• Quality of data • Compatibility of outcomes • Population differences • Healthcare differences	Not used.	
	German colonoscopy screening programme (2003- 2007)[17]	• Compatibility of outcomes • Population differences • Healthcare differences	Used. Large sample sizes (> 4 million people) and geographical similarity.[17, 42-44]	• Adjust the calibration targets to address differences in definitions • Down-weight the target in the calibration process • Validate (to predicted sensitivity of screening FIT20 and FS for HRA)
Undiagnosed CRC by age and sex	German colonoscopy screening programme (2003- 2007)[17]	• Population differences • Healthcare differences	USED. The only data source available	• Adjust the calibration target to reflect the difference in national incidence rates • Adjust the calibration target to the timing of the events • Down-weight the target in the calibration process • Validate (to predicted sensitivity of screening FIT20 and FS for CRC)
Undiagnosed CRC by Duke stages by age and sex	German multi-centre cohort study in 2003-2010[20]	• Quality of data • Compatibility of the outcomes • Population differences • Sample differences • Healthcare differences	USED. The only data source available	• Compare the definitions of CRC stages • Down-weight the target in the calibration process
CRC incidence by age and sex (in	2005 Cancer Registration Statistics/	• Quality of data • Age of data	USED.	• Up-weight the target in the calibration process

absence of screening programme)	census data for England (2005)[33, 45]		Representative for all-England population	
	Oxford regional data 2005	• Population differences	Not used.	
CRC incidence by age, sex, and Duke stages (in absence of screening programme)	UKACR National Colorectal dataset 1996-2004.	• Quality of data (Missing stage at diagnosis for 43% of patients) Age of data	USED. The only national data source available	• Split the missing stage at diagnosis between stages C and D to fit the survival predictions • Down-weight the target in the calibration process

The Legend: CRC – colorectal cancer, FIT - faecal immunochemical test, FS – flexible sigmoidoscopy, HRA – high-risk adenoma, LRA – low-risk adenoma, RCT – randomised controlled trial