UNIVERSITAT AUTÒNOMA DE BARCELONA

DOCTORAL THESIS

Calibration in Cost-Effectiveness modeling (research diary)

Author:
David GÓMEZ

Supervisors:
Dr. Josep Lluis ARCOS
Dr. Mireia DÍAZ

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Artificial Intelligence Research Institute (IIIA)
Escola d'Enginyeria de la UAB

July 4, 2022

Contents

Chapter 1

Background

1.1 Cost-Effectiveness Models in Healthcare

1.1.1 Objective

Compare different strategies for detection/treatment of a disease, from health and economic point of view.

1.1.2 Methodology

Simulation model to mimic the strategies and compare the outputs for each strategy to determine which strategies are worth considering. A special "strategy" called the natural history describes the progression of the disease without any planned interventions, and it is used to calibrate some the inputs that will be used in the rest of strategies (see section 3).

1.1.3 Types of model

Decision trees, markov model, microsimulation, ... depending on the needs of the domain and the degree of detail and granularity required.

1.1.4 Inputs

Parameters extracted from the scientific literature, studies, expert opinions, assumptions, ... An interesting intermediate input are the transition matrices that show the probabilities of transitioning between health states.

1.1.5 Outputs

For each strategy:

- Effectiveness measure (e.g. Quality-Adjusted Life Years, QALYs)
- Cost measure (e.g. euros, €)
- Other general measures of interest: incidence, mortality, ...
- Other domain-dependent measures: e.g. number of hysterectomies, number of high-grade lesions, ...

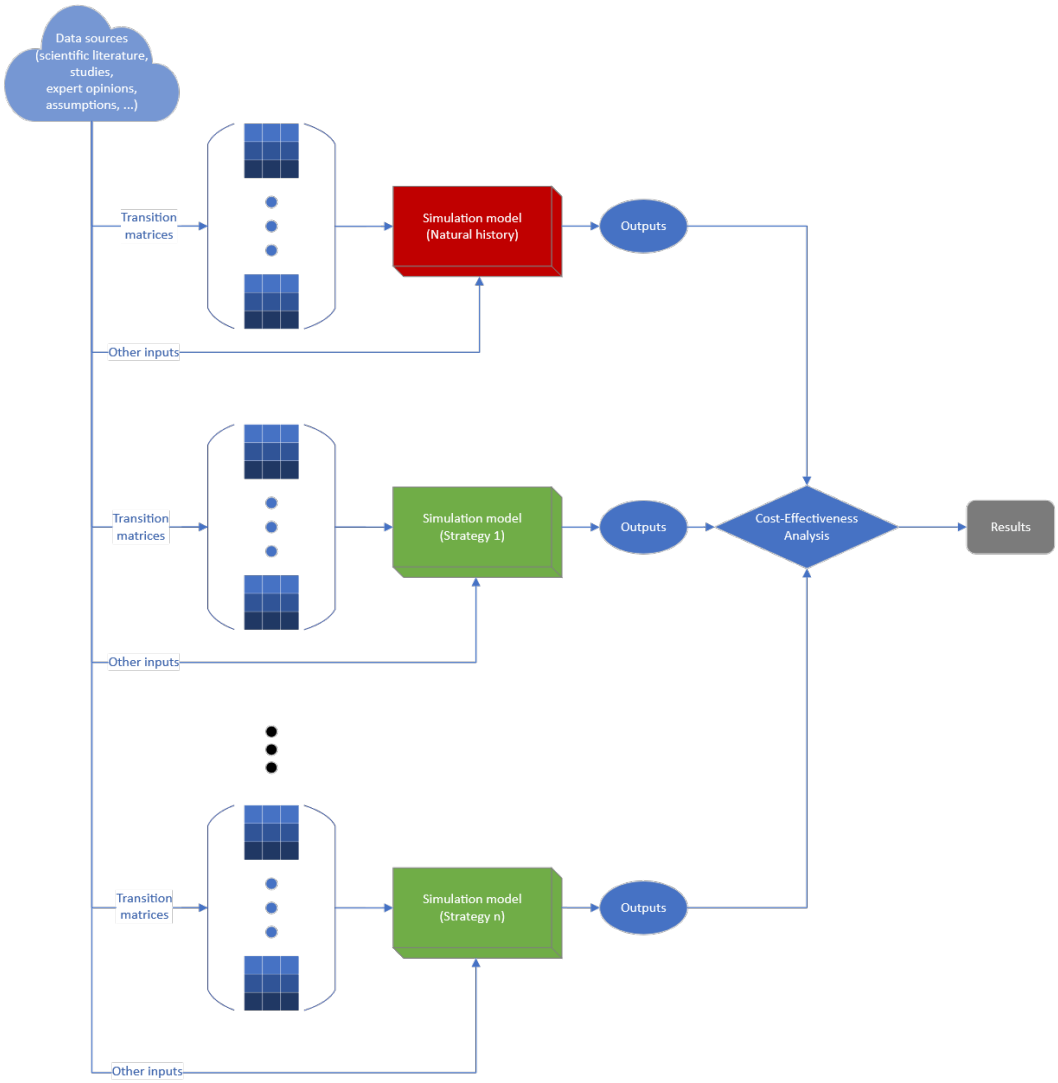

FIGURE 1.1: Overview of the cost-effectiveness analysis.

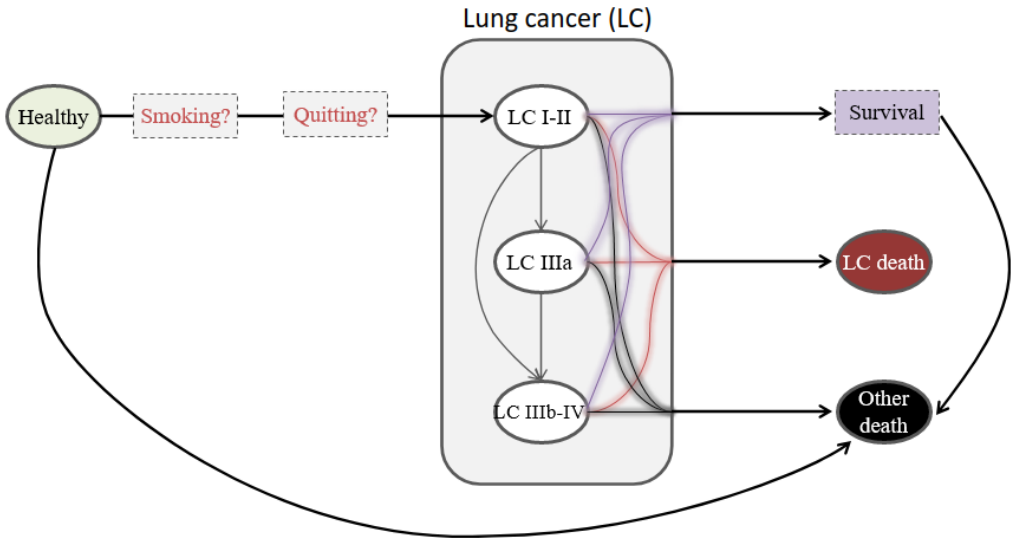

FIGURE 1.2: Markov state diagram of the lung cancer model.

Chapter 2

Cost-Effectiveness Analysis Methodology

2.1 Calibration

Before starting the base analysis, we calibrate the transition matrices in the natural history by slightly modifying the original probabilities so that the output of our model (e.g. incidence, mortality, ...) fits an observed value based on evidence. These calibrated probabilities can then be used by the rest of the strategies in the base analysis. See Calibration Workflow for more details.

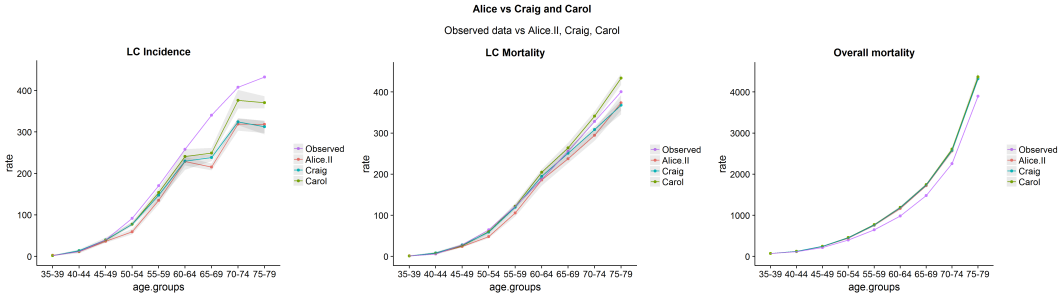

FIGURE 2.1: Example of three calibration curves for incidence, lung cancer mortality and mortality from other causes, along with the expected outcome.

2.2 Base analysis

Each strategy is plotted in the Cost and Effectiveness axes and the efficiency curve shows the strategies that are cost-effective, the rest are dominated by them and they are not considered cost-effective.

We can compare each strategy in relation to another calculating the Incremental Cost-Effectiveness Ratio as:

$$ICER = \frac{\Delta C}{\Delta E} = \frac{C_2 - C_1}{E_2 - E_1}$$

If the ICER is below the Willingness-To-Pay (WTP) threshold (i.e. the maximum amount of money a country/region is willing to pay per additional QALY) the second strategy is more cost-effective than the first. If the ICER is greater than the WTP the strategy's benefits are not considered cost-effective (i.e. the increased

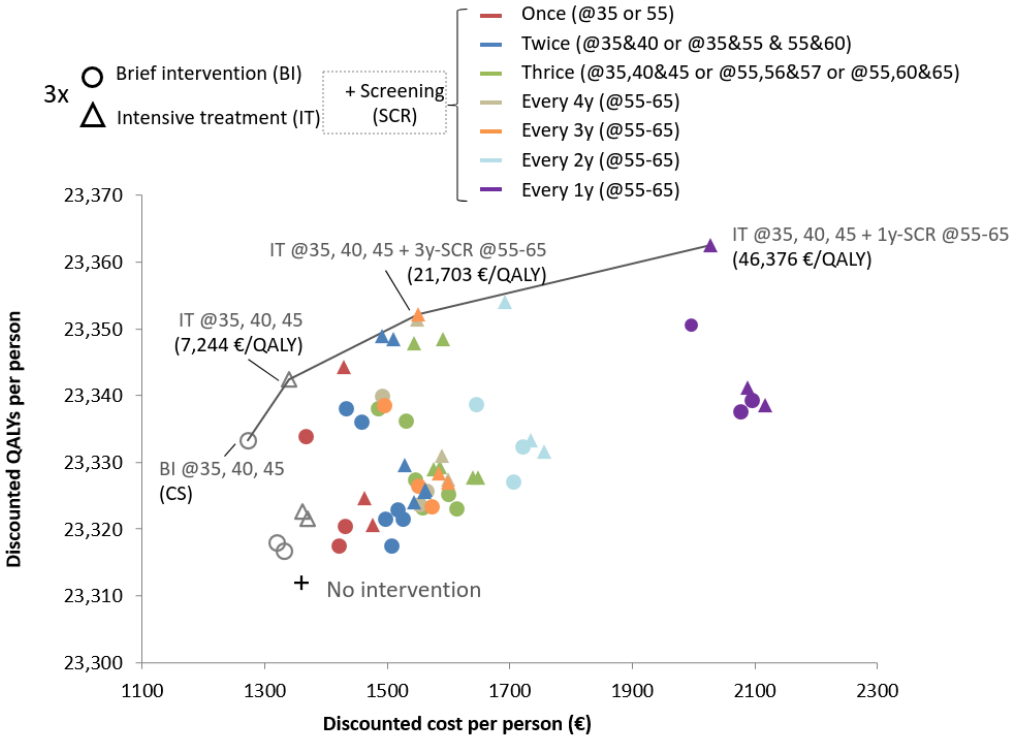

FIGURE 2.2: Example of an efficiency curve showing the cost (€) and effectiveness (QALYs) of each of the simulated strategies. The strategies on the curve represent the cost-effective strategies, those below it are not usually considered since they are always dominated by one of those in the curve.

health benefit does not justify the increment of cost). Negative ICERs imply that one strategy dominates the other one.

2.3 Sensitivity analysis

Once the base analysis is performed we evaluate the uncertainty of the used parameters to check the robustness of the results. We modify the values of the parameters of interest to see how they affect the output of the model.

2.3.1 Deterministic Sensitivity Analysis (DSA)

A sweep is performed over a range (e.g. +-15% of the base value) for the parameters of interest, to see how the ICER changes and whether the cost-effectiveness decision is different.

2.3.2 Probabilistic Sensitivity Analysis (PSA)

Each parameter of interest is modeled as a probabilistic distribution (e.g. Beta for probabilities, Gamma/Lognormal for costs, ...) with the base value as the mean and a standard deviation dependent on the amount of uncertainty. We sample from these distributions (univariate or multivariate) to run a number of random simulations

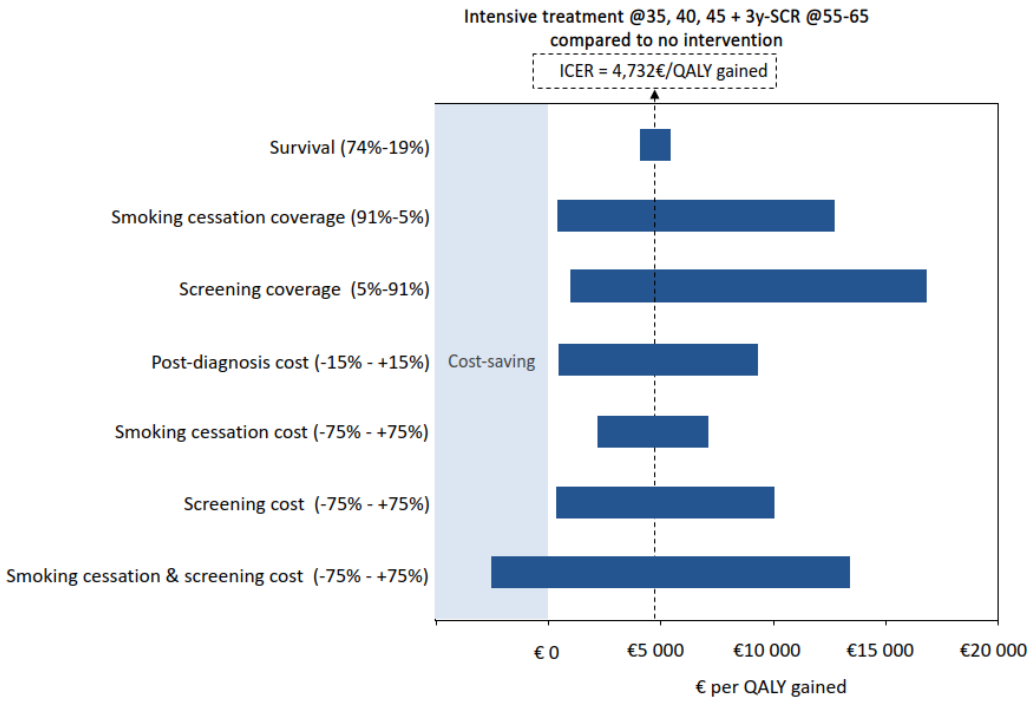

FIGURE 2.3: Example of a tornado diagram showing the impact on the ICER between two particular strategies when changing one single parameter over a predefined range.

to check the percentage of simulations that show a cost-effective result (i.e. the percentage of simulations below the line ICER=WTP, see figure 2.4).

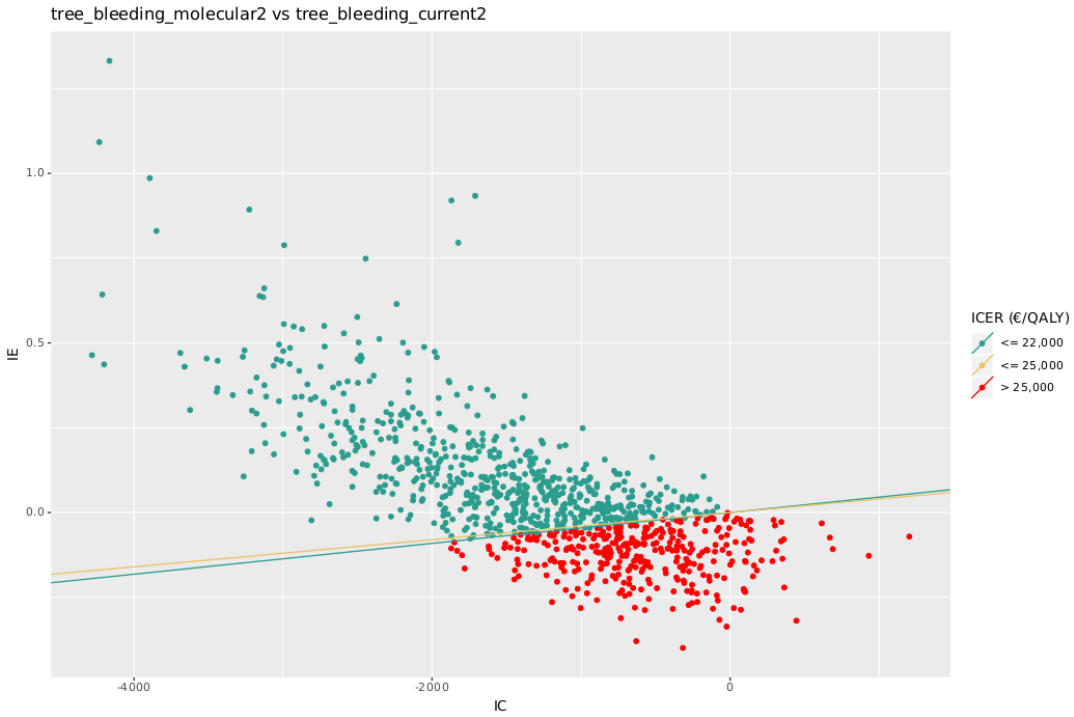

FIGURE 2.4: Example of a scatterplot of a PSA, showing results from 1,000 random iterations and how they relate to the cost-effectiveness threshold (WTP). Green and red points represent cost-effective and non-cost-effective simulations respectively.

Chapter 3

Calibration workflow

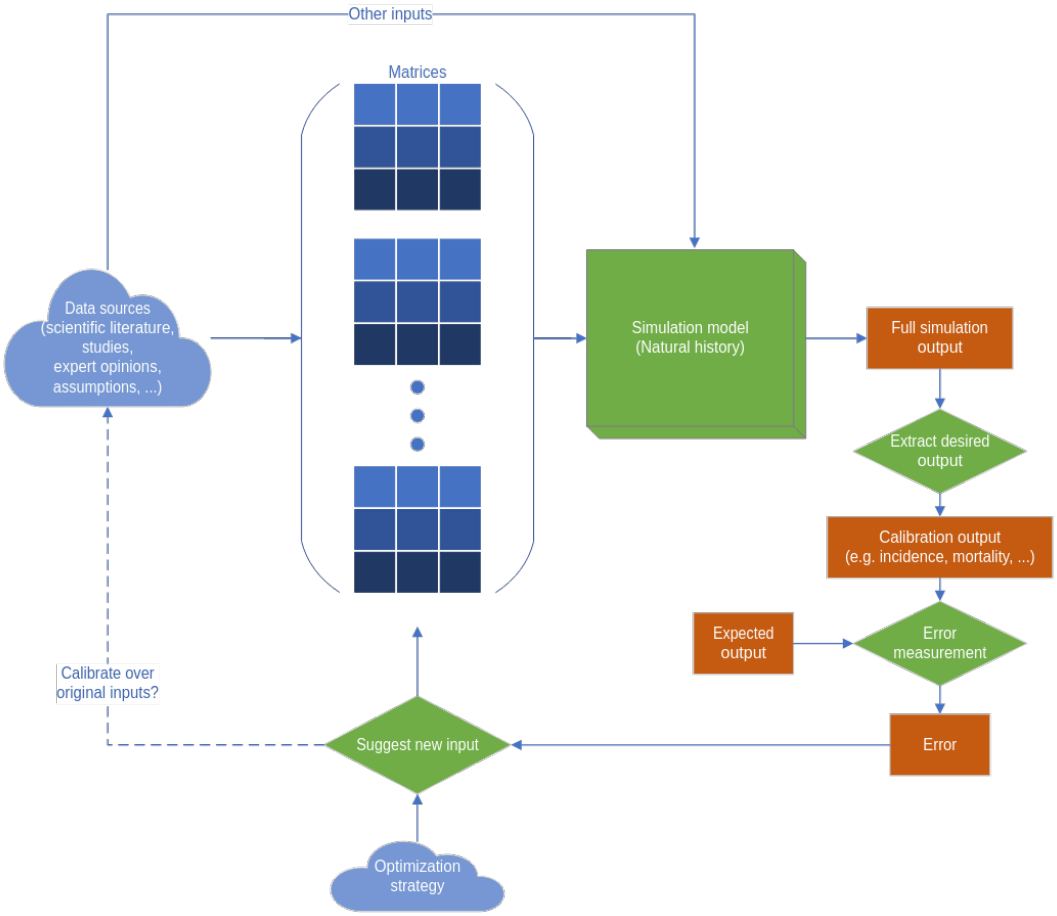

FIGURE 3.1: Conceptual summary of the calibration procedure.

Chapter 4

Research proposal

Cost-effectiveness models are used to evaluate and compare different medical strategies (e.g. strategies to detect cancer), in terms of both health and cost, and to help decision makers determine the optimal allocation of medical resources. These models have several inputs that describe the environment, the disease and the strategies used (inputs like probabilities, sensitivities and specificities for medical procedures, costs, utility values, ...) and can have several outputs as needed for the analysis (e.g. average life expectancy, average cost, incidence of the disease, mortality, ...).

The stated goal of the research project is to find efficient ways to calibrate these cost-effectiveness models by changing some input parameters so a particular output (usually incidence or mortality) matches a given value found in the scientific literature. In summary, **we can frame the problem as the optimization of a black box function computing (for example) the euclidean distance between the simulated output and a theoretical value**.

In some of our models we find that classical optimization methods need many simulations to converge to a good solution, demanding a lot of time and computing resources. Also, in many cases we are able to provide information about the statistical distribution of some of the inputs that could help in the optimization process. For these reasons I believe Bayesian Optimization (BO) might be a good alternative to adapt to our particular modeling needs.

Some preliminary (ongoing) experiments, using both bayesian and classical methods, are available in a jupyter notebook in `https://github.com/david-gomez-guillen/phd`, using the hypermapper [Nardi et al., 2019] python library for BO and scipy for classical optimization methods. This first set of tests focus on the performance of different optimization methods on well-known analytical functions that are easy to compute. Once acquainted with BO usage, the next step would be to repeat the tests on the actual cost-effectiveness models.

The conclusions and challenges found in the tests so far include:

1. BO using Gaussian Processes might be too slow for some of our more lightweight simulation models, classical methods converge faster even if they need more function evaluations. It should not be a problem for more computationally expensive simulation models.

2. BO tests performed converge to worse optima compared to classical optimization methods (e.g. Nelder-Mead, BFGS). It might be due to code issues: either implementation problems or an incorrect usage of the library.

3. Gaussian processes regression used in BO becomes more expensive for each iteration due to having to calculate the inverse of a matrix that grows with the number of observations. Regression becomes very slow to compute after a number of evaluations, especially for high-dimensional inputs (depending

on the available hardware) [Das, Roy, and Sambasivan, 2015][Terry and Choe, 2021].

4. Regular bayesian methodology (updating iteratively the model after each observation) is more difficult to parallelize than classical methods, though some alternatives exist [Daulton, Balandat, and Bakshy, 2021][Wang et al., 2016].

5. BO in these tests performed with priors for the optimum [Souza et al., 2020] (implemented in hypermapper: `https://github.com/luinardi/hypermapper/wiki/prior-injection`) do not seem to converge faster. It might be due to library issues too: optimum priors implementation is labeled as experimental in the documentation.

6. When optimizing the actual cost-effectiveness models in future tests it would be interesting to add constraints for the inputs [Ungredda and Branke, 2021][Gardner et al., 2014]. E.g: one input parameter being greater than another.

Chapter 5

Thoughts & ideas

5.1 Calibration over original inputs

In the standard calibration procedure we change the transition probabilities in the matrices, assuming they are independent and enforcing constraints in the error calculation step. One alternative would be to calibrate over those relevant original inputs (i.e. those found in the literature, studies, expert opinions, ...) and then calculate the transition probabilities (and possibly other inputs of the model) from them. With this method we preserve and inject the knowledge we have about the domain into the model (that is, how to calculate the probabilities from the scientific sources, assumptions, implicit restrictions, ...).

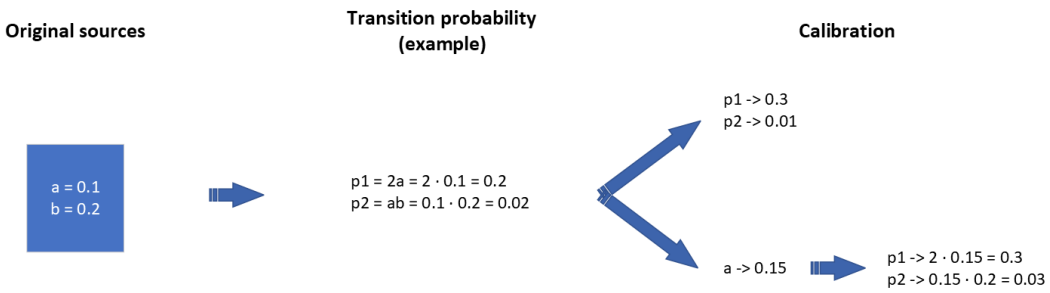

FIGURE 5.1: Calibration procedures: directly over the probabilities ("Calibration", top) and over the original inputs ("Calibration", bottom).

Since we calibrate to account for the uncertainty in the natural history and the probabilities in the matrices are calculated from parameters, we could classify the uncertainty sources in two: uncertainty associated to the inputs and uncertainty associated to the calculation of the probability.

5.1.1 Parameter uncertainty

If we are sure about the calculation of the transition probability, like a well-known relationship (e.g. Bayes formula), the remaining sources of uncertainty are the inputs themselves. In this case we can set up the optimizer to change these inputs, recalculate the probabilities and run the model.

For interpretation and sanity check purposes we can check both the changed input value and the newly-calculated transition probability.

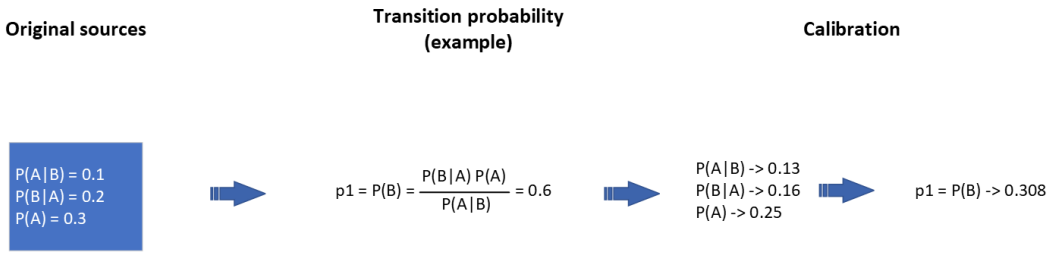

FIGURE 5.2: If we are certain of the relationship between the probability and the inputs, we can calibrate over the inputs and we will get the calibrated probability by preserving the link with the original inputs.

5.1.2 Calculation uncertainty

Beyond the inputs, the calculation itself might be uncertain as well, for example by making a very rough approximation or assuming a probabilistic distribution with a poor fit. In this case we can insert an error term or scaling factor in the formula to account for the misspecification, with a neutral initial value (0 if error, 1 if factor). Then, the calibration could include these error/scaling terms in the set of parameters to be optimized.

For interpretation and sanity check purposes, besides the probabilities themselves as usual, we can check the error term/scaling factor. If they are too different from the initial values we might conclude that the calculation is not trustworthy and we might have to review our assumptions. Also, if the calculation is a very rough estimate another alternative would be to reject the calculation itself and optimize the probability value as usual.

5.1.3 Comments

Pros

- Preserving the link and implicit knowledge between the original sources and the calculated probabilities
- Preserving relationships between probabilities and (some kinds of) constraints

Cons

- Overcomplicating the calibration procedure in simple cases
- Relationships might not be too complicated

5.2 Input/probabilities dependencies using graph theory

5.2.1 Comments

Pros

- Might help the calibration procedure by exploiting domain knowledge

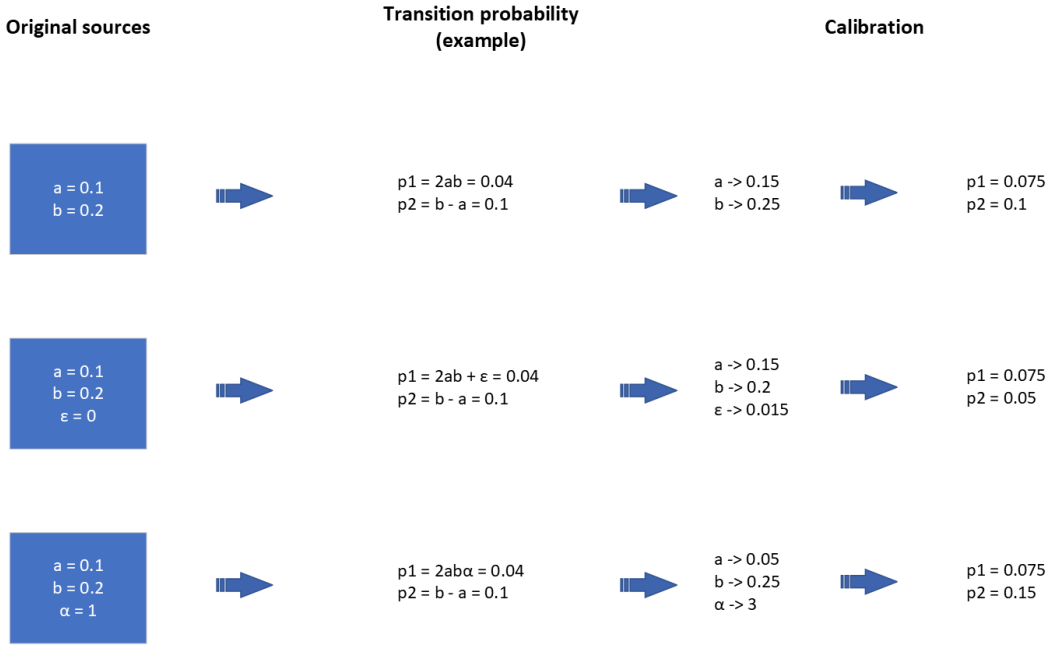

FIGURE 5.3: To account for the possibility of a misspecified calculation we can add additional terms to a formula. We can see calibration with no formula flexibility (top row), an additive error term for p1 (middle row) or a multiplicative factor for p1 (bottom row).

Cons

- Overcomplicating the calibration procedure in simple cases
- Relationships might not be too complicated
- Too vague at this point

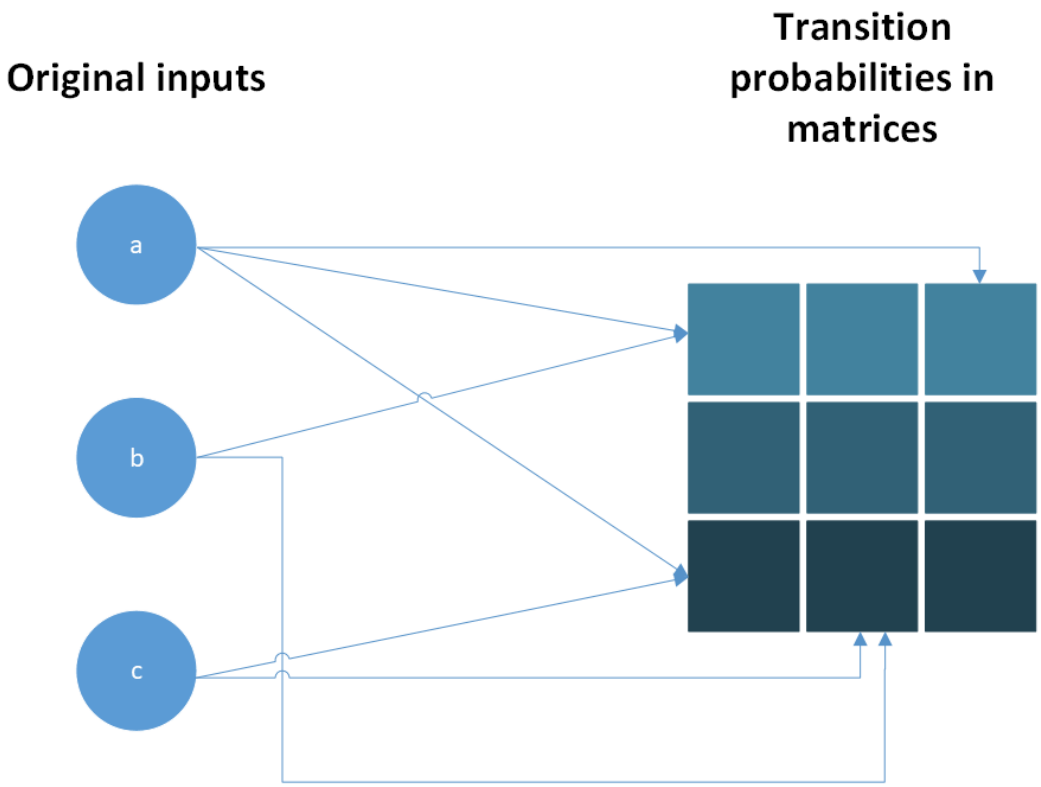

FIGURE 5.4: We could use graph theory to help optimization considering the dependencies between inputs and the transition probabilities in the matrices.

Chapter 6

Tests performed

6.1 Test model: lung cancer

- Total of 9 matrices (one per age group: 35-39, 40-44, . . . , 75-79).
- 7 health states: healthy, stages I-II, stage IIIa, stage IIIb, LC survival, death from LC, death from other causes → 7x7 matrices. Sometimes the LC survival state is excluded from the calibration resulting in 6x6 matrices.
- Matrices represent monthly steps in the simulation. Since they are applied for 5-year groups, each matrix is used in 5 * 12 = 60 iterations in the model.
- A simplified calibration can be performed without running the model, only the matrices are used. This is a fast approximation since we are not considering some factors of the full model (e.g. prevalence of smoking):
  - LC survival state is excluded → 6x6 matrices
  - From the 6x6=36 probabilities per matrix, only 11 probabilities are allowed to change. The rest are either constant (zeroes, ones) or one minus the sum of the rest of the row.
  - The error measurement is a weighted sum of the absolute differences of the LC incidence, LC mortality and mortality from other causes. The weights are 0.45, 0.45 and 0.10 respectively.

6.1.1 Simplified calibration, 1 matrix

- Source file: models/lung/calibration_wrapper.R (N_MATRICES set to 1)
- Only the first age group is being calibrated (35-39): 1x11 = 11 parameters.

Algorithm	Initial matrix	Nelder-Mead	Particle swarm	Bayesian
Error	1.1545799674960	0.6633085653748	0.66298515	0.6629851533965
Time (s)	-	0.76	22.97	114.89
Evaluations	-	252	10100	21

6.1.2 Simplified calibration, 2 matrices

- Source file: models/lung/calibration_wrapper.R (N_MATRICES set to 2)
- The first and second age groups are being calibrated (35-39 and 40-44): 2x11 = 22 parameters.

Algorithm	Initial matrix	Nelder-Mead	Particle swarm	Bayesian
Error	1.6495138536869	0.7333381543348456	0.72801101	0.7287116136105645
Time (s)	-	4.44	24.46	421.69
Evaluations	-	2092	10100	70

6.1.3 Simplified calibration, all (9) matrices

- Source file: models/lung/calibration_wrapper.R (N_MATRICES set to 9)
- All age groups are being calibrated: 9x11 = 99 parameters.
- Standard bayesian optimization takes too much time and the process was aborted before completion. Other strategies could be attempted: calibrate matrices sequentially, restrict number of parameters, optimize gaussian process regression (see section 4), ...

Algorithm	Initial matrix	Nelder-Mead	Particle swarm	Bayesian
Error	6.6842251473203	4.0210661197016	4.3594607	<Aborted>
Time (s)	-	57.58	35.64	-
Evaluations	-	19800	9407	-

List of Abbreviations

QALY	**Q**uality **A**djusted **L**ife **Y**ear
CEA	**C**ost **E**ffectiveness **A**nalysis
WTP	**W**illingness **T**o **P**ay

Bibliography

Das, Sourish, Sasanka Roy, and Rajiv Sambasivan (2015). "Fast Gaussian Process Regression for Big Data". In: *CoRR* abs/1509.05142. arXiv: 1509.05142. URL: http://arxiv.org/abs/1509.05142.

Daulton, Samuel, Maximilian Balandat, and Eytan Bakshy (2021). *Parallel Bayesian Optimization of Multiple Noisy Objectives with Expected Hypervolume Improvement*. DOI: 10.48550/ARXIV.2105.08195. URL: https://arxiv.org/abs/2105.08195.

Gardner, Jacob et al. (2014). "Bayesian Optimization with Inequality Constraints". In: *Proceedings of the 31st International Conference on Machine Learning*. Ed. by Eric P. Xing and Tony Jebara. Vol. 32. Proceedings of Machine Learning Research 2. Bejing, China: PMLR, pp. 937–945. URL: https://proceedings.mlr.press/v32/gardner14.html.

Nardi, Luigi et al. (2019). "HyperMapper: a Practical Design Space Exploration Framework". In: *27th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, MASCOTS 2019, Rennes, France, October 21-25, 2019*. IEEE Computer Society, pp. 425–426. DOI: 10.1109/MASCOTS.2019.00053. URL: https://doi.org/10.1109/MASCOTS.2019.00053.

Souza, Artur et al. (2020). *Bayesian Optimization with a Prior for the Optimum*. DOI: 10.48550/ARXIV.2006.14608. URL: https://arxiv.org/abs/2006.14608.

Terry, Nick and Youngjun Choe (Aug. 2021). "Splitting Gaussian processes for computationally-efficien regression". In: *PLOS ONE* 16.8, pp. 1–17. DOI: 10.1371/journal.pone.0256470. URL: https://doi.org/10.1371/journal.pone.0256470.

Ungredda, Juan and Juergen Branke (2021). *Bayesian Optimisation for Constrained Problems*. DOI: 10.48550/ARXIV.2105.13245. URL: https://arxiv.org/abs/2105.13245.

Wang, Jialei et al. (2016). *Parallel Bayesian Global Optimization of Expensive Functions*. DOI: 10.48550/ARXIV.1602.05149. URL: https://arxiv.org/abs/1602.05149.