Improving Cycle Corrections in Discrete Time Markov Models: A Gaussian Quadrature Approach

Tushar Srivastava, MSc*[1], Mark Strong, PhD†[1], Matthew D Stevenson, PhD‡[1], and Peter J Dodd, PhD§[1],**

[1]School of Health and Related Research (ScHARR), University of Sheffield, UK

**Peter J Dodd, School of Health and Related Research (ScHARR), University of Sheffield, Regent Court, 30 Regent Street, Sheffield, S1 4DA, UK (p.j.dodd@sheffield.ac.uk)

July 27, 2020

Abstract

Introduction: Discrete-time Markov models are widely used within health economic modelling. Analyses usually associate costs and health outcomes with health states and calculate totals for each decision option over some timeframe. Frequently, a correction method (e.g. half-cycle correction) is applied to unadjusted model outputs to yield an approximation to an assumed underlying continuous-time Markov model. In this study, we introduce a novel approximation method based on Gaussian Quadrature (GQ).

Methods: We exploited analytical results for time-homogeneous Markov chains to derive a new GQ-based approximation, which is applied to an unadjusted discrete-time model output. The GQ method approximates a continuous-time Markov model result by approximating a correction matrix, formulated as an integral, using a weighted sum of integrand values at specified points. GQ approximations can be made arbitrarily accurate by increasing 'order' of the approximation. We compared the first five orders of GQ approximation with four existing cycle correction methods (half-cycle correction, trapezoidal and Simpson's 1/3 and 3/8 rules) across 100,000 randomly generated input parameter-sets.

Results: We show that first-order GQ method is identical to half-cycle correction method, which is itself equivalent to trapezoidal method. The second-order GQ is identical to Simpson's 1/3 method. The third, fourth and fifth order GQ methods are novel in this context and provide increasingly accurate approximations to the output of the continuous-time model. In our simulation study, fifth-order GQ method outperformed other existing methods in over 99.8% of simulations. Of the existing methods, Simpson's 1/3 rule performed the best.

Conclusion: Our novel GQ-based approximation outperforms other cycle correction methods for time-homogeneous models. The method is easy to implement, and R code and an Excel workbook are provided as supplementary materials.

*t.srivastava@sheffield.ac.uk (ORCID ID:0000-0002-5961-9348)

†m.strong@sheffield.ac.uk (ORCID ID:0000-0003-1486-8233)

‡m.d.stevenson@sheffield.ac.uk (ORCID ID:0000-0002-3099-9877)

§p.j.dodd@sheffield.ac.uk (ORCID ID:0000-0001-5825-9347)

1 Introduction

Discrete-time state-transition Markov models are common in healthcare decision analytical modelling. A discrete-time Markov model assumes that transitions between health states in a disease or treatment process occur at fixed intervals [1,2], but in reality, transitions between health states can usually occur at any point in time. A *continuous*-time Markov model is therefore a more natural representation for most disease and treatment processes. [1–3] Forcing transitions to occur only at discrete time steps leads to biased predictions of cumulative outcomes (e.g. cumulative costs, or cumulative utilities), and it is common for some kind of correction method to be applied. A correction method (e.g. the half-cycle correction [HCC]), applied to the cumulative outcomes of an unadjusted discrete-time state-transition Markov model, will result in model output that is closer to that which would have been observed in the continuous time analogue. [1–4]

There are various cycle correction method that exist, [5] the most common being the half-cycle correction method. [3, 6] This method assumes that, on average, transitions occur *halfway* through each time step, rather than at the beginning or end of each cycle, and simply subtracts and adds half of the outcome in each health state from the first and last cycle, respectively (note that we use the terms 'cycle' and 'time step' interchangeably.) [3, 6] Other methods from the numerical integration literature have also been used in decision analytical models for cycle correction. [5] These methods include the trapezoidal rule, Simpson's 1/3 rule, and Simpson's 3/8 rule, which all apply corrections at each cycle, not only at the first and last cycle. [5] Recently, Elbasha and Chhatwal (2016) reported that Simpson's 1/3 method has the better accuracy compared to trapezoidal and Simpson 3/8 method. [5]

In this paper, we propose a new approach for cycle correction based on Gaussian quadrature (GQ). Gaussian quadrature is a method for approximating the integral of a function using a weighted sum of evalutions of the function at carefully chosen points. [7] We show that a correction factor, which adjusts a discrete-time model so that it matches its 'true' continuous time analogue, can be expressed as an integral, and this integral can be approximated using Gaussian quadrature.

The structure of the paper is as follows: in the next section we describe the relationship between a discrete-time and an assumed underlying continuous-time Markov model. We then review four existing correction methods: the half-cycle correction method, the trapezoidal rule, Simpson's 1/3 rule, and Simpson's 3/8 rule. Our novel correction method based on Gaussian Quadrature is then introduced, along with a simulation study to demonstrate its performance relative to existing methods. We conclude with a short discussion and conclusions. Supplementary material to implement the GQ approximation in Microsoft Excel and in R can be downloaded from the journal website.

2 Discrete- and continuous-time Markov models

2.1 The discrete-time Markov model

We start by introducing a simple illustrative cohort Markov model with three health states: 'well', 'unwell' and 'dead'. At each time step, well patients can remain well, move to the unwell state, or die. Patients in the unwell state can recover to the well state, remain unwell, or die. Dead is an absorbing state of the model. The probabilities of transition between states are described by transition matrix,

$$\mathbf{M} = \begin{array}{c c} & \begin{array}{ccc} \text{well} & \text{unwell} & \text{dead} \end{array} \\ \begin{array}{c} \text{well} \\ \text{unwell} \\ \text{dead} \end{array} & \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.05 & 0.65 & 0.3 \\ 0 & 0 & 1 \end{bmatrix} \end{array}, \tag{1}$$

As the population cohort is assumed to be closed, each row of transition matrix sums to 1.

We assume that the proportion of individuals in each health state at time zero is given by the row vector $\mathbf{s}_0^\top = (1, 0, 0)$, i.e. all individuals start in the 'well' state. The proportion of individuals in each health state at time step 1 is then given by the matrix product of $\mathbf{s}_0^\top$ by $\mathbf{M}$,

$$\mathbf{s}_1^\top = \mathbf{s}_0^\top \mathbf{M}. \tag{2}$$

The proportion of individuals in each health state at time step t is found by repeated *matrix* multiplication of $\mathbf{M}$,

$$\mathbf{s}_t^\top = \mathbf{s}_0^\top \mathbf{M}^t. \tag{3}$$

Note that $\mathbf{M}^t$ is not equal to the matrix obtained by raising the elements of $\mathbf{M}$ to the t-th power.

In our example, health states are associated with the following per-time step costs ($\mathbf{c}$) and utilities ($\mathbf{u}$),

$$\mathbf{c} = \begin{bmatrix} 5 \\ 100 \\ 0 \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} 0.95 \\ 0.6 \\ 0 \end{bmatrix}. \tag{4}$$

The total cost is determined by summing the costs associated with each time step over some predefined number of time steps (the 'time horizon'), and similarly for total utility. In our example, the time horizon is 100 time steps. For brevity, we present methods and results for costs only. All arguments and expressions apply equally to utilities, simply by replacing $\mathbf{c}$ with $\mathbf{u}$.

The cost accrued at time step t, denoted $C(t)$, is given by

$$C(t) = \mathbf{s}_t^\top \mathbf{c} = \mathbf{s}_0^\top \mathbf{M}^t \mathbf{c}, \tag{5}$$

and the total cost over the period of time from $t = 0$ to $t = N$ is given by

$$T_c^{DT} = \sum_{t=0}^{N-1} C(t) = \sum_{t=0}^{N-1} \left(\mathbf{s}_0^\top \mathbf{M}^t \mathbf{c} \right) = \mathbf{s}_0^\top \left(\sum_{t=0}^{N-1} \mathbf{M}^t \right) \mathbf{c}. \tag{6}$$

The superscript DT on the total costs T_c denotes that it is derived from the discrete-time model. In the discrete-time model as formulated in Equation (6), costs are assumed to occur at the *start* of each time period, hence t runs from 0 to $N - 1$. If we assumed instead that outcomes occurred at the *end* of each time period, total costs would be given by $\mathbf{s}_0^\top \left(\sum_{t=1}^{N} \mathbf{M}^t \right) \mathbf{c}$, where t now runs from 1 to N.

See Figure 1(a) for a graphical illustration of how total costs are computed in a hypothetical discrete-time model where costs are assumed to occur at the start of each time period.

Discrete-time Markov models are attractive due to their computational simplicity. However, in the health economic evaluation context in which disease processes are modelled, a discrete-time model will only approximate what is, in reality, a continuous process. We now introduce the continuous-time Markov model.

2.2 The continuous-time Markov model

Markov models can also be defined in continuous time via the Kolmogorov Forward Equation, a differential equation of the form

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathbf{s}_t^\top = \mathbf{s}_t^\top \mathbf{R}, \tag{7}$$

where $\mathbf{s}_t^\top$ is a vector containing the proportion of individuals in each health states at time t, and $\mathbf{R}$ is a matrix of transition *rates* (sometimes also called the generator or intensity matrix). [8,9] When $\mathbf{R}$ is constant (i.e. where rates are time-homogeneous), the evolution from the initial state to time t can be written in terms of a matrix exponential as

$$\mathbf{s}_t^\top = \mathbf{s}_0^\top \exp(\mathbf{R}t), \tag{8}$$

where $\mathbf{s}_0^\top$ is the row vector containing the proportion of individuals in each health state at time zero.

Every time-homogeneous continuous-time Markov model gives rise to a discrete-time Markov model that, for unit time step, has transition probability matrix $\mathbf{M} = \exp(\mathbf{R})$, where, again, exp() denotes the matrix exponential. For our model, the rate matrix that corresponds to transition probability matrix $\mathbf{M}$ is

$$\mathbf{R} = \begin{bmatrix} -0.368 & 0.299 & 0.069 \\ 0.075 & -0.422 & 0.368 \\ 0 & 0 & 0 \end{bmatrix}. \tag{9}$$

We can verify that $\mathbf{M} = \exp(\mathbf{R})$ using the `expm()` matrix exponential function in the R package `expm`, and there are formulae for computing the matrix exponential by hand in low dimensions. [10] A discrete-event simulation for this continuous-time Markov chain could be implemented using 0.299 as the hazard of progressing from 'well' to 'unwell', 0.069 as the hazard of progressing from 'well' to 'dead', and so on.

In transition rate matrix $\mathbf{R}$, rows sum to zero. This corresponds to the rows of $\mathbf{M}$ summing to 1, and indicates a closed cohort. The third row of zeroes corresponds to the absorbing state 'dead', which is not left once entered, and where there is no flow between states.

The cost at time t, denoted $C(t)$, is given by

$$C(t) = \mathbf{s}_t^\top \mathbf{c} = \mathbf{s}_0^\top \exp(\mathbf{R}t)\mathbf{c}, \tag{10}$$

and the total cost over the time period from $t = 0$ to $t = N$ is given by the integral

$$T_c^{CT} = \int_0^N C(t)\,\mathrm{d}t = \int_0^N \mathbf{s}_t^\top \mathbf{c}\,\mathrm{d}t = \int_0^N \mathbf{s}_0^\top \exp(\mathbf{R}t)\mathbf{c}\,\mathrm{d}t,$$

where the superscript CT on the total costs T_c denotes that it is derived from the continuous-time model. We can compute this integral analytically via the basic rules of integration, giving

$$T_c^{CT} = \mathbf{s}_0^\top \{\exp(\mathbf{R}N) - \mathbb{I}\}\mathbf{R}^{-1}\mathbf{c}. \tag{11}$$

This is illustrated in Figure 1(b).

We note here that because the rate matrix, $\mathbf{R}$, has rows that sum to zero, it will not have a full set of linearly independent columns, and will therefore not be invertible. However, equation (11) can still be evaluated if we replace the usual inverse with the *Moore-Penrose generalised inverse*. [11] This is implemented as the `ginv()` function in the `MASS` package in R.

3 Existing methods to correct a discrete-time model output

As we have noted, a discrete-time model is likely to be an approximation to a 'true' underlying continuous-time process. A discrete-time model assumes that transitions occur *at the time step*. No events occur *between* time steps because in a discrete time model, there is no time between steps; time is not continuous. In reality, in most settings, transitions between disease states can occur at any point in time. Therefore, an adjustment must be made to the output of a discrete-time model to address the bias that arises from the approximation.

3.1 Half-Cycle correction method

The 'Half-Cycle' correction (HCC) method is commonly used in health economic modelling. HCC assumes that the transition of the patient from one health state to another occurs midway between time steps, rather than at the time step. It is computed by subtracting and adding half of the outcome in the first and last cycle of the discrete-time model (equation 6), respectively, giving

$$\begin{aligned} T_c^{HCC} &= T_c^{DT} - \frac{1}{2}\mathbf{s}_0^\top \mathbf{M}^0 \mathbf{c} + \frac{1}{2}\mathbf{s}_0^\top \mathbf{M}^N \mathbf{c} \\ &= \sum_{t=0}^{N-1} \mathbf{s}_0^\top \mathbf{M}^t \mathbf{c} - \frac{1}{2}\mathbf{s}_0^\top \mathbf{M}^0 \mathbf{c} + \frac{1}{2}\mathbf{s}_0^\top \mathbf{M}^N \mathbf{c}. \end{aligned} \tag{12}$$

This is illustrated in Figure 1(c).

3.2 Trapezoidal method

This numerical integration technique, also commonly employed in health economic modelling, approximates an integral by dividing the range over which the function is integrated into trapeziums, and summing the areas of those trapeziums, [12] as illustrated in Figure 1(d).

Given two adjacent cycles, t and $t+1$, with corresponding costs $\mathbf{s}_0^\top \mathbf{M}^t \mathbf{c}$ and $\mathbf{s}_0^\top \mathbf{M}^{t+1} \mathbf{c}$, the area of the trapezium over this interval is given by

$$\mathbf{A}_t = \frac{1}{2}(\mathbf{s}_0^\top \mathbf{M}^t \mathbf{c} + \mathbf{s}_0^\top \mathbf{M}^{t+1} \mathbf{c}). \tag{13}$$

The total cost over the time horizon N is then given by:

$$\begin{aligned} T_c^{Trap} &= \sum_{t=0}^{N-1} \mathbf{A}_t = \sum_{t=0}^{N-1} \frac{1}{2}(\mathbf{s}_0^\top \mathbf{M}^t \mathbf{c} + \mathbf{s}_0^\top \mathbf{M}^{t+1} \mathbf{c}) \\ &= \sum_{t=1}^{N-1} \mathbf{s}_0^\top \mathbf{M}^t \mathbf{c} + \frac{1}{2}(\mathbf{s}_0^\top \mathbf{M}^0 \mathbf{c} + x_0 \mathbf{M}^N \mathbf{c}) \\ &= \sum_{t=0}^{N-1} \mathbf{s}_0^\top \mathbf{M}^t \mathbf{c} - \frac{1}{2}\mathbf{s}_0^\top \mathbf{M}^0 \mathbf{c} + \frac{1}{2}\mathbf{s}_0^\top \mathbf{M}^N \mathbf{c} = T_c^{HCC}, \end{aligned} \tag{14}$$

which, as shown in equation (14), is exactly equal to that given by half cycle correction method. The trapezoidal method is illustrated in Figure 1(d).

3.3 Simpson's composite methods

Simpson's composite methods uses a higher order polynomial function to connect adjacent points on the cost curve, rather than using a straight-line segment as used in the trapezoidal method. There are two commonly used types of Simpson's rule: a quadratic approximation-based rule (Simpson's 1/3), and a cubic approximation-based rule (Simpson's 3/8) [13, 14]

3.3.1 Simpson's 1/3 rule

Given the costs computed for three adjacent cycles $\mathbf{s}_0^\top \mathbf{M}^t \mathbf{c}$, $\mathbf{s}_0^\top \mathbf{M}^{t+1} \mathbf{c}$ and $\mathbf{s}_0^\top \mathbf{M}^{t+2} \mathbf{c}$, Simpson's rule computes the area under a section of quadratic that passes through the three points. This is repeated then for the next three cycles $\mathbf{s}_0^\top \mathbf{M}^{t+2} \mathbf{c}$, $\mathbf{s}_0^\top \mathbf{M}^{t+3} \mathbf{c}$ and $\mathbf{s}_0^\top \mathbf{M}^{t+4} \mathbf{c}$, and so on.

The total cost can be shown to be

$$T_c^{S_{1/3}} = \frac{1}{3}\mathbf{s}_0^\top \mathbf{M}^0 \mathbf{c} + \frac{2}{3}\sum_{\substack{t>0 \\ \text{even}}}^{N-2} \mathbf{s}_0^\top \mathbf{M}^t \mathbf{c} + \frac{4}{3}\sum_{\substack{t=1 \\ \text{odd}}}^{N-1} \mathbf{s}_0^\top \mathbf{M}^t \mathbf{c} + \frac{1}{3}\mathbf{s}_0^\top \mathbf{M}^N \mathbf{c} \tag{15}$$

The Simpson's 1/3 correction can be obtained by multiplying the outcomes by 1/3 in the first and last cycle and by 4/3 if the cycle number is odd and by 2/3 if the cycle number is even. Simpson's 1/3 rule is illustrated in Figure 1(e). Note that in this method the time horizon N must be even.

3.3.2 Simpson's 3/8 rule

Similar to Simpson's 1/3 rule, Simpson's 3/8 rule uses sections of polynomial, but this time, cubic polynomial fitted to four adjacent points.

The total cost can be shown to be

$$T_c^{S_{3/8}} = \frac{3}{8}\mathbf{s}_0^\top \mathbf{M}^0 \mathbf{c} + \frac{6}{8}\sum_{\substack{t>0 \\ \bmod 3=0}}^{N-3} \mathbf{s}_0^\top \mathbf{M}^t \mathbf{c} + \frac{9}{8}\sum_{\substack{t=1 \\ \bmod 3\neq 0}}^{N-1} \mathbf{s}_0^\top \mathbf{M}^t \mathbf{c} + \frac{3}{8}\mathbf{s}_0^\top \mathbf{M}^N \mathbf{c} \tag{16}$$

The Simpson's 3/8 rule correction can be obtained by by multiplying the outcomes by 3/8 in the first and last cycle, by 6/8 if the cycle number is a multiple of 3 and by 9/8 otherwise. Simpson's 3/8 rule is illustrated in Figure 1(f). Note that in this method the time horizon N must be an multiple of three.

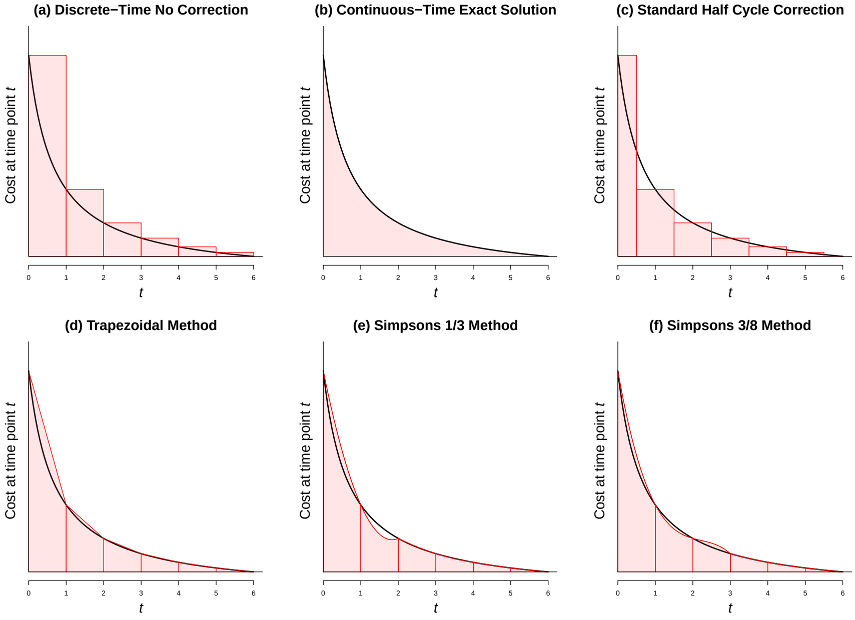

Figure 1: *Graphical illustration of the (a) discrete- and (b) continuous-time models, and four correction methods (c-f) for the discrete-time model. The areas shaded in red represent total costs over a time horizon from $t = 0$ to $t = 6$.*

4 New cycle correction method based on Gaussian quadrature

Our approach is to consider the difference between the total costs from the discrete- and continuous- time models, and attempt to find an approximation to this difference. Firstly, we derive an analytic expression for the total costs for a discrete-time model (6) by recognising that it is the sum of a finite geometric progression. For a *scalar*, α, the sum of a finite geometric progression is given by

$$\sum_{n=0}^{N} \alpha^n = (\alpha^{N+1} - 1)(\alpha - 1)^{-1}, \tag{17}$$

for any $\alpha \neq 1$. The matrix analogue for (17) is given by

$$\sum_{n=0}^{N} \mathbf{M}^n = (\mathbf{M}^{N+1} - \mathbb{I})(\mathbf{M} - \mathbb{I})^{-1}, \tag{18}$$

where $\mathbb{I}$ is the identity matrix (in this case of dimension 3), and hence we can write the discrete-time model as

$$T_c^{DT} = \left(\sum_{t=0}^{N-1} \mathbf{s}_0^\top \mathbf{M}^t \mathbf{c} \right) = \mathbf{s}_0^\top \left(\sum_{t=0}^{N-1} \mathbf{M}^t \right) \mathbf{c} = \mathbf{s}_0^\top (\mathbf{M}^N - \mathbb{I})(\mathbf{M} - \mathbb{I})^{-1} \mathbf{c}. \tag{19}$$

As an aside, this formula is only valid when $(\mathbf{M} - \mathbb{I})$ is invertible, which will *not* be the case if $\mathbf{M}$ is a transition matrix. This is because $\mathbf{M} - \mathbb{I}$ will have rows that sum to zero, and therefore will not have a full set of linearly independent columns. However, we can get around this problem, as long as the costs and utilities are zero for absorbing states ('dead' in our example), by replacing the usual matrix inverse with the Moore-Penrose generalised inverse (see Appendix for proof). [11, 15]

We now consider how we could adjust the discrete-time model so that the cumulative sum of outcomes equal those of the continuous-time model. Recall that the rate and transition matrices are linked via $\mathbf{M} = \exp(\mathbf{R})$ and hence $\mathbf{M}^N = \exp(\mathbf{R}N)$. We can therefore get to the continuous-time model cumulative sum of costs

$$T_c^{CT} = \mathbf{s}_0^\top \{\exp(\mathbf{R}N) - \mathbb{I}\} \mathbf{R}^{-1} \mathbf{c}, \tag{20}$$

from the discrete-time model cumulative sum of costs

$$T_c^{DT} = \mathbf{s}_0^\top (\mathbf{M}^N - \mathbb{I})(\mathbf{M} - \mathbb{I})^{-1} \mathbf{c}, \tag{21}$$

by inserting a correction factor of $(\mathbf{M} - \mathbb{I}) \log(\mathbf{M})^{-1}$ before $\mathbf{c}$ in the expression (21), as follows

$$\begin{aligned} &\mathbf{s}_0^\top (\mathbf{M}^N - \mathbb{I})(\mathbf{M} - \mathbb{I})^{-1} (\mathbf{M} - \mathbb{I}) \log(\mathbf{M})^{-1} \mathbf{c} \\ &= \mathbf{s}_0^\top (\mathbf{M}^N - \mathbb{I}) \log(\mathbf{M})^{-1} \mathbf{c} \\ &= \mathbf{s}_0^\top (\mathbf{M}^N - \mathbb{I}) \mathbf{R}^{-1} \mathbf{c} \\ &= \mathbf{s}_0^\top \{\exp(\mathbf{R}N) - \mathbb{I}\} \mathbf{R}^{-1} \mathbf{c} \\ &= T_c^{CT}. \end{aligned} \tag{22}$$

Recall that the Markov trace is the matrix that records the health state occupancy over time. It has the same number of columns as there are health states, and the same number of rows as there are time cycles. In our example, the matrix is of size 100 × 3. Summing down the columns of the Markov trace gives us $\sum_{t=0}^{N-1} \left(\mathbf{s}_0^\top \mathbf{M}^t \right)$, which we showed above can be written $\mathbf{s}_0^\top (\mathbf{M}^N - \mathbb{I})(\mathbf{M} - \mathbb{I})^{-1}$. So, in practice, we simply need to post-multiply the sum of the columns of the Markov trace from our discrete time model by the correction factor $(\mathbf{M} - \mathbb{I}) \log(\mathbf{M})^{-1}$ before we then multiply by the cost vector $\mathbf{c}$, i.e. replacing $\left(\sum_{t=0}^{N-1} \mathbf{s}_0^\top \mathbf{M}^t \right) \mathbf{c}$ with $\left(\sum_{t=0}^{N-1} \mathbf{s}_0^\top \mathbf{M}^t \right) (\mathbf{M} - \mathbb{I}) \log(\mathbf{M})^{-1} \mathbf{c}$.

We now turn to the question of how we find $(\mathbf{M} - \mathbb{I}) \log(\mathbf{M})^{-1}$. If we are using a coding language (R, Matlab, Python, etc) then we can evaluate $(\mathbf{M} - \mathbb{I}) \log(\mathbf{M})^{-1}$ directly using in-built matrix algebra functions. In R the correction factor would be `(M-I) %*% ginv(logm(M))`, where `ginv()` is the Moore-Penrose inverse function from the `MASS` package, `logm()` is the matrix log function from the `expm` package, and `I` is defined as the identity matrix of dimension n matching the dimensionality of $\mathbf{M}$, via `I <- diag(nrow(M))`. If the sum of the columns of the Markov trace is denoted `x` in R, and our cost vector by `c`, then we can recover the continuous-time model total costs via `x %*% (M-I) %*% ginv(logm(M)) %*% c`.

This is trivial in R, but unfortunately, the matrix logarithm is very difficult to compute in a typical spreadsheet application such as Microsoft Excel, and this problem motivates the final step: the Gaussian quadrature approximation.

Firstly, we define $\mathbf{Z}$ to be the *inverse* of the correction factor (we will invert it back again later),

$$\begin{aligned} \mathbf{Z} &= \{(\mathbf{M} - \mathbb{I}) \log(\mathbf{M})^{-1}\}^{-1} \\ &= \log(\mathbf{M})(\mathbf{M} - \mathbb{I})^{-1}. \end{aligned}$$

Next, we note that $\mathbf{Z}$ has an integral representation as

$$\mathbf{Z} = \log(\mathbf{M})(\mathbf{M} - \mathbb{I})^{-1} = \int_0^1 \{(\mathbf{M} - \mathbb{I})x + \mathbb{I}\}^{-1} \mathrm{d}x. \tag{23}$$

The proof of this is somewhat involved, and we refer interested readers to Wouk (1965) [16].

Finally, we develop an approximation to the integral representation of $\mathbf{Z}$ using Gaussian quadrature, in such a way that we can compute the approximation in a spreadsheet. Gaussian quadrature is a method for approximating the integral of a function by a sum of carefully chosen polynomials. In practice, this involves evaluating the function to be integrated at a set of pre-specified points, and summing the values with pre-specified weights. The points and weights can be found in standard tables. The number of terms in the summation is called the *order* of the GQ approximation, with higher orders giving better approximations.

Several Gaussian quadrature 'rules' (i.e. methods for determining the points and weights) exist, including *Gauss-Legendre* quadrature, which approximates the integral representation of $\mathbf{Z}$ in Equation 23 as

$$\int_0^1 f(x)\mathrm{d}x = \frac{1}{2}\sum_{i=1}^{n} w_i f\left(\frac{1+x_i}{2}\right), \tag{24}$$

where $f(x) = \{(\mathbf{M} - \mathbb{I})x + \mathbb{I}\}^{-1}$, the function to be integrated.

The locations $\{x_i\}$ and weights $\{w_i\}$ for first 5 orders (values of n) are given in Table 1.

Order	Locations $\{\mathbf{s}_i^\top\}$	Weights $\{w_i\}$
1	0	2
2	$\pm\sqrt{\frac{1}{3}}$	1
3	0	$\frac{8}{9}$
	$\pm\sqrt{\frac{3}{5}}$	$\frac{5}{9}$
4	$\pm\sqrt{\frac{3}{7} - \frac{2}{7}\sqrt{\frac{6}{5}}}$	$\frac{18+\sqrt{30}}{36}$
	$\pm\sqrt{\frac{3}{7} + \frac{2}{7}\sqrt{\frac{6}{5}}}$	$\frac{18-\sqrt{30}}{36}$
5	0	$\frac{128}{225}$
	$\pm\frac{1}{3}\sqrt{5 - 2\sqrt{\frac{10}{7}}}$	$\frac{322+13\sqrt{70}}{900}$
	$\pm\frac{1}{3}\sqrt{5 + 2\sqrt{\frac{10}{7}}}$	$\frac{322-13\sqrt{70}}{900}$

Table 1: *Locations and weights for the first 5 orders of Gaussian-Legendre quadrature.*

Plugging these values into Equation 24, gives the following approximations for $\mathbf{Z}$ at orders $n = 1$ and 2:

$$\mathbf{Z} \approx \hat{\mathbf{Z}}_1 = \frac{1}{2} \times 2 \times f\left(\frac{1+0}{2}\right) = \{(\mathbf{M} - \mathbb{I})\frac{1}{2} + \mathbb{I}\}^{-1} = 2(\mathbf{M} + \mathbb{I})^{-1}. \tag{25}$$

$$\mathbf{Z} \approx \hat{\mathbf{Z}}_2 = \frac{1}{2}\left(\frac{1+\sqrt{\frac{1}{3}}}{2}\mathbf{M} + \frac{1-\sqrt{\frac{1}{3}}}{2}\mathbb{I}\right)^{-1} + \frac{1}{2}\left(\frac{1-\sqrt{\frac{1}{3}}}{2}\mathbf{M} + \frac{1+\sqrt{\frac{1}{3}}}{2}\mathbb{I}\right)^{-1}. \tag{26}$$

Orders 3, 4 and 5 can be calculated similarly by evaluating $f(x)$ at the locations in Table 1, and summing with the relevant weights. Fully-worked formulae are also in the supplementary material.

To summarise, we run the unadjusted discrete-time Markov model to generate the Markov trace, sum down the columns of the Markov trace, post-multiply this by the inverse of $\hat{\mathbf{Z}}_n$ (where n is the chosen order for the GQ correction), and finally post-multiply the result by the cost vector $\mathbf{c}$.

The first order GQ approximation is equivalent to the trapezoidal method (which, as we have shown, is equivalent to the half-cycle correction) This can be shown as follows:

$$\begin{aligned} T_c^{GC(1)} &= \sum_{t=0}^{N-1} \mathbf{s}_0^\top \mathbf{M}^t \left\{2(\mathbf{M}+\mathbb{I})^{-1}\right\}^{-1} \mathbf{c} = \sum_{t=0}^{N-1} \mathbf{s}_0^\top \mathbf{M}^t \left\{\frac{1}{2}(\mathbf{M}+\mathbb{I})\right\} \mathbf{c} \\ &= \sum_{t=0}^{N-1} \frac{1}{2}\left(\mathbf{s}_0^\top \mathbf{M}^t \mathbf{c} + \mathbf{s}_0^\top \mathbf{M}^{t+1} \mathbf{c}\right) = T_c^{Trap}. \end{aligned} \tag{27}$$

The second order GQ method can be shown to be identical to Simpson's 1/3 method, but the third order GQ method is novel and should result in a better approximation than either the half cycle or Simpson's 1/3 methods. Higher order GQ methods will results in better approximations still.

4.1 Discounting

Normally, analyses in health economics employ discounting. For discrete-time models, we multiply costs accrued at time step t by a factor $1/(1+\rho)^t$, where ρ is the per-time-step discount rate. This is equivalent to replacing the transition probability matrix $\mathbf{M}$ with $\mathbf{M}^* = \mathbf{M}/(1+\rho)$,

$$T_c^{DT}(\text{discounted}) = \sum_{t=0}^{N-1} \left(\mathbf{s}_0^\top \mathbf{M}^t \mathbf{c} \times \frac{1}{(1+\rho)^t}\right) = \mathbf{s}_0^\top \left(\sum_{t=0}^{N-1} \mathbf{M}^{*t}\right) \mathbf{c}. \tag{28}$$

For a continuous-time model, we would add the natural logarithm of the discount factor $\rho^* = \log\{1/(1+\rho)\}$ to the diagonal elements of $\mathbf{R}$, giving $\mathbf{R}^* = \mathbf{R} + \rho^*\mathbb{I}$, and therefore

$$T_c^{CT}(\text{discounted}) = \mathbf{s}_0^\top \{\exp(\mathbf{R}^* N) - \mathbb{I}\} \mathbf{R}^{*-1} \mathbf{c}. \tag{29}$$

All the approximation methods we have described in this section hold equally when discounting is employed and we will not consider this further, other than to note that $(\mathbf{M}^* - \mathbb{I})$ and $\mathbf{R}^*$ are both invertible, and we therefore do not need to use the Moore-Penrose generalised inverse when outcomes are discounted.

5 Simulation Study

We first undertook a deterministic analysis in which we assumed that the transition matrix $\mathbf{M}$, cost vector $\mathbf{c}$, utility vector $\mathbf{u}$ and starting health state proportion vector $\mathbf{s}_0^\top$ took, with certainty, their values as presented above.

We then performed a probabilistic sensitivity analysis (PSA) to explore the impact of uncertainty on input parameters of the model. [17] In this analysis, transition probabilities, costs and utilities were each assigned a probability distribution. For costs, we assumed lognormal distributions, for utilities, we assumed beta distributions, and for the rates in matrix $\mathbf{R}$ we assumed exponential distributions (Table 2). The transition probability matrix $\mathbf{M}$ was derived by taking the matrix exponential of the rate matrix R. All parameters were considered independent. See supplemental Figure S4 for histograms of these distributions.

Model Inputs	Mean Value	Standard Deviation	Distribution Used in PSA
Well State Cost	£5	£1	Log Normal (1.6,0.198)
Unwell State Cost	£100	£20	Log Normal (4.61,0.198)
Dead State Cost	£0	£0	Assumed zero with certainty
Well State Utility	0.95	0.19	Beta (0.3,0.016)
Unwell State Utility	0.6	0.12	Beta (9.4,6.27)
Dead State Utility	0	0	Assumed zero with certainty
r_{12} (Well to unwell rate)	0.299	0.299	Exp(1/0.299)
r_{13} (Well to dead rate)	0.069	0.069	Exp(1/0.069)
r_{21} (Unwell to well rate)	0.075	0.075	Exp(1/0.075)
r_{23} (Unwell to dead rate)	0.368	0.368	Exp(1/0.368)
r_{31} (Dead to well rate)	0	0	Assumed zero with certainty
r_{31} (Dead to unwell rate)	0	0	Assumed zero with certainty
r_{11} (Well to well rate)			$-(r_{12} + r_{13})$
r_{22} (Unwell to unwell rate)			$-(r_{21} + r_{23})$
r_{33} (Dead to dead rate)			$-(r_{31} + r_{32})$

Table 2: *Model parameters with mean values, standard deviation and probabilistic distribution used for PSA. Rate r_{ij} is the entry in the i^{th} row and j^{th} column of the rate matrix* $\mathbf{R}$. *Abbreviations - PSA: Probabilistic Sensitivity Analysis*

To ensure good coverage of parameter space, a Latin hypercube design was used to draw 100,000 parameter sets. For each parameter set, we computed the total costs and total QALYs for the discrete-time model, continuous-time model, and the six approximation methods: (i) half-cycle = trapezoidal = 1st order GQ; (ii) Simpson's 1/3 rule = 2nd order GQ; (iii) Simpson's 3/8 rule; (iv) 3rd order GQ; (v) 4th order GQ; and (vi) 5th order GQ. In each case, we computed the absolute relative error (i.e. the absolute value of the difference between the approximation and the continuous model output, divided by the continuous model output) for costs and for QALYs.

To understand the impact of cycle length on approximations, we also performed scenario analysis with varying cycle lengths from 1 week to 52 weeks.

All the analyses were performed in both Microsoft Excel and in R, and code can be found in the supplementary material.

6 Results

Results are presented in Table 3. In the deterministic analysis, the 3rd order GQ method outperformed existing methods, with higher order GQ approximations further improving the approximation. Simpson's 1/3 method (which is identical to the 2nd order GQ method) was the best performing of the existing methods. For total costs in our example, the half-cycle correction method (equivalent to the trapezoidal and 1st order GQ methods) gave a worse approximation than no correction. In addition, the difference in net-monetray benefit (NMB) between our proposed method and the half-cycle method is over £400 which implies that the method employed for continuity correction could alter decision-making.

Results with discounting are presented in supplementary Table S1

Correction Method	Costs	Absolute Error	QALYs	Absolute Error	NMB
Continuous time model	£228.7622	-	4.27397	-	-
5th order GQ method	£228.7622	1.86e-07	4.27397	6.85e-07	-£0.0877
4th order GQ method	£228.7621	3.33e-07	4.27397	7.03e-07	-£0.0900
3rd order GQ method	£228.7604	7.73e-06	4.27396	1.55e-06	-£0.1972
2nd order GQ method	£228.6821	3.51e-04	4.27386	2.54e-05	-£3.1738
Simpson's 3/8 method	£228.5886	7.59e-04	4.27374	5.29e-05	-£6.6033
1st order GQ method	£226.4474	1.01e-02	4.28816	3.32e-03	£427.9517
No correction performed	£228.9474	8.09e-04	4.76316	1.14e-01	£14,675.4517

Table 3: *Comparative analysis result per person based on different methods. Absolute error is presented in increasing order of error. 5th order GQ method has the absolute minimum error for both costs and QALYs and closest to continuous model. 1st order GQ is same as HCC and trapezoidal and 2nd order GQ is same as Simpson 1/3. NMB is with respect to continuous time model with willingness-to-pay of £30,000. Abbreviations: GQ: Gaussian Quadrature; HCC: Half-Cycle Correction; NMB: Net Monetary Benefit; QALY: Quality Adjusted Life-Years*

The PSA results are presented in Figure 2. The box and whisker plot presents the absolute relative errors on the log scale for each correction method, compared against the continuous-time model output. The 5th order GQ method has the smallest median absolute relative errors, followed in increasing order of error by the 4th order GQ, 3rd order GQ and then Simpson's 1/3 method.

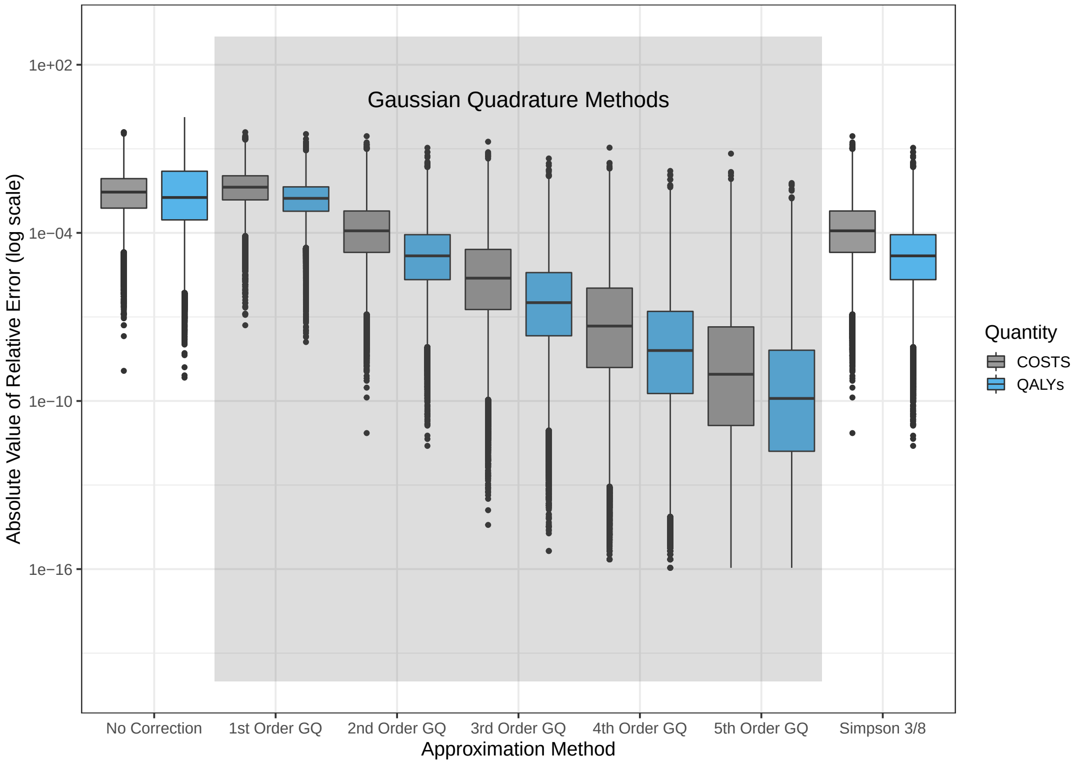

Figure 2: *Probabilistic Sensitivity Analysis Result: Absolute value of the relative errors for cost and QALY in 100,000 PSA runs between the continuous time model and discrete time model with different cycle correction methods.1st order GQ is same as HCC and trapezoidal; 2nd order GQ is same as Simpson 1/3. Costs and QALYs outcomes corrected with 5th order GQ method has the minimum error. Abbreviations: GQ: Gaussian Quadrature; HCC: Half-Cycle Correction; QALY: Quality Adjusted Life-Years*

A rankogram in Figure 3 shows that in 100,000 PSA runs, the 5th order GQ approximation method performed best in the 99.8% of cases (i.e. it was not ranked first in 0.2% of runs). Whereas, Simpson's 1/3 rule (the best of the existing methods) ranked first in only 0.001% runs.

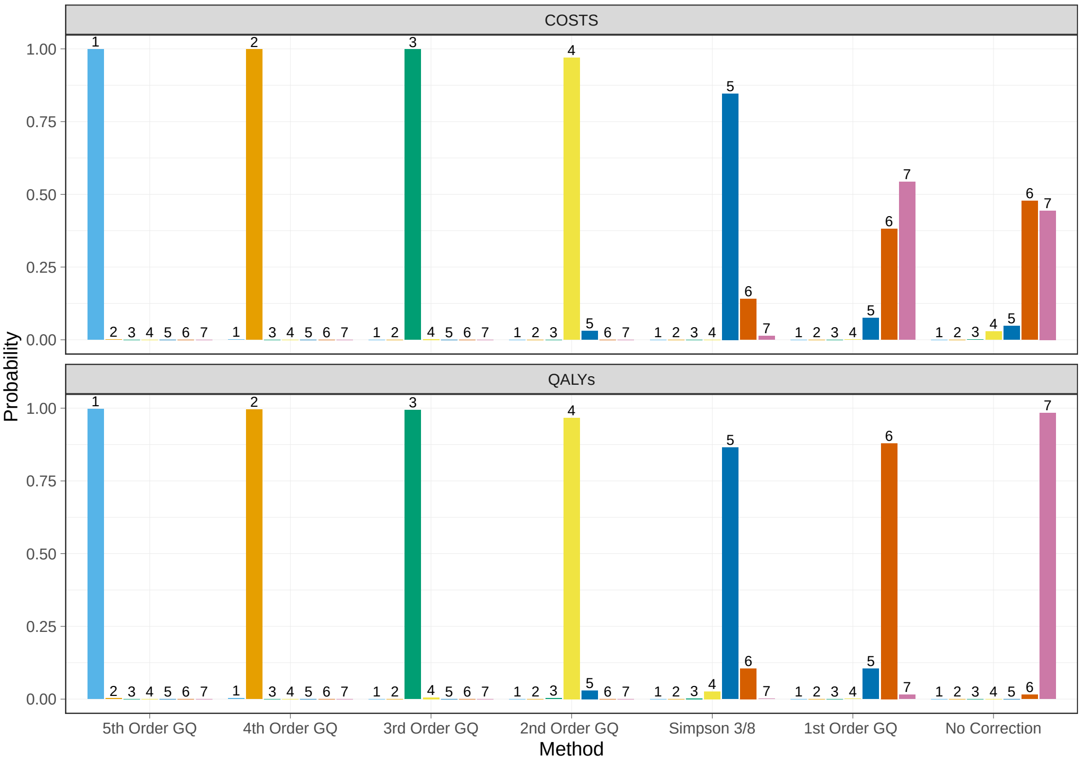

Figure 3: *Rankogram bar plot showing ranks for accuracy (absolute value of relative error) of each approximation method in 100,000 PSA runs; where rank 1 represents minimum error and rank 7 maximum error. 1st order GQ is same as HCC and trapezoidal ; 2nd order GQ is same as Simpson 1/3. Abbreviations: GQ: Gaussian Quadrature; HCC: Half-Cycle Correction; PSA: Probabilistic Sensitivity Analysis; QALY: Quality Adjusted Life-Years*

When we examined the parameters of the runs in which the 5th order GQ approximation was not best, we discovered two patterns. Firstly, the 5th order GQ approximation suffered when the transition probabilities for well to well, and from unwell to unwell were both close to 1 (i.e. then transitions between states were rare, and $\mathbf{M}$ was close to the identity matrix). In these cases, all methods performed well, with the best being the 4th rather than the 5th order GQ method. However, the differences between the 4th and 5th order approximations were tiny, with absolute relative errors of the order 10^{-14}, and therefore unlikely to be of any consequence. These runs form the vast majority of 5th-order GQ 'failures', and appear as a peak in the density plots shown in supplementary Figures S2 and S3. Secondly, the 5th order GQ approximation did less well when the probabilities for 'forward' transitions (i.e. from well to unwell, from well to dead, and from unwell to dead) were high, resulting in 100% mortality within a few cycles. Indeed, in this case, *all* of the correction methods performed poorly, with the uncorrected output being closest to the continuous-time model result. There were very few failures of this type ($< 0.01\%$ of all runs, and these failures are therefore not visible in supplementary Figures S2 and S3.)

The impact of cycle length on the quality of the approximations is presented in Figure 4. With a shorter cycle length, the absolute error between continuous- and discrete-time model using different approximation methods decreases as expected. The 5th order GQ method performs best overall.

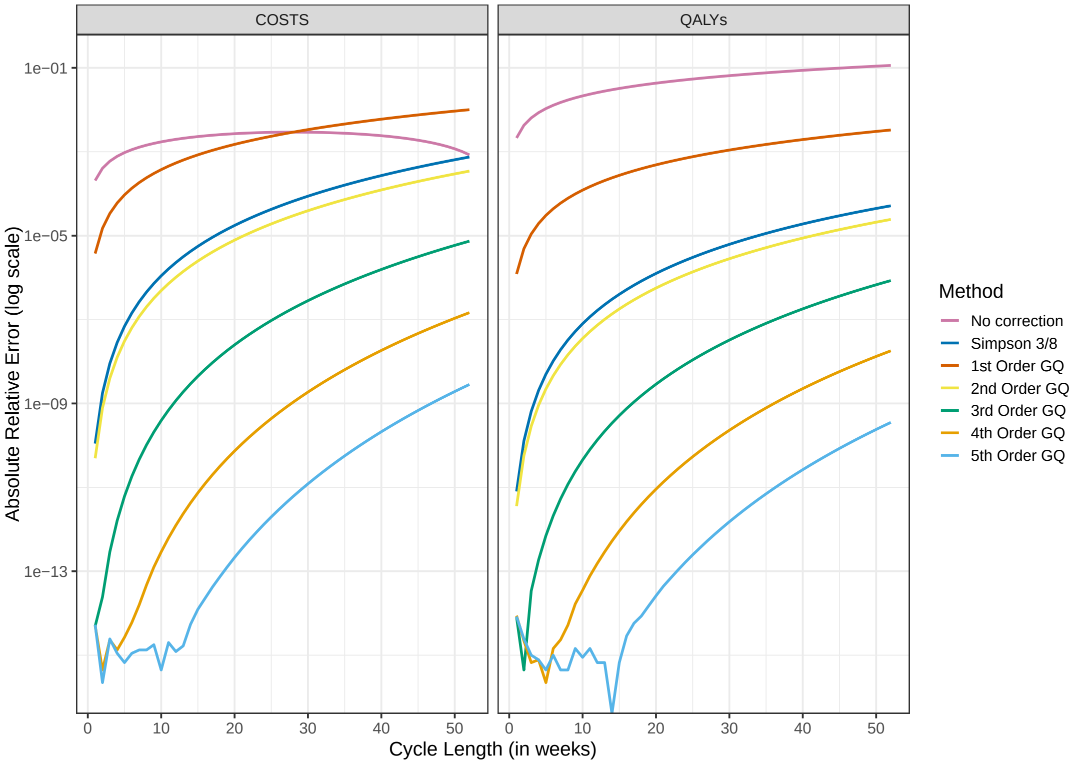

Figure 4: *Impact of cycle length on absolute error between continuous- and discrete-time model using different approximation methods.1st order GQ is same as HCC and trapezoidal and 2nd order GQ is same as Simpson 1/3. Most of the time 5th order GQ method has lowest error compared to the other methods. Abbreviations: GQ: Gaussian Quadrature; HCC: Half-Cycle Correction; QALY: Quality Adjusted Life-Years*

The impact of time horizon on the quality of approximation is presented in supplementary Figure S1 and similar results were observed. In each case, the 5th order GQ method had the lowest error.

7 Discussion

7.1 Main results

In our study, we found that the 3rd, 4th and 5th order GQ methods all outperformed existing cycle correction approximation methods, with higher orders giving better approximations. We chose the 5th as the highest order GQ to implement, but the adjusted discrete-time model output can be made arbitrarily close to the continuous-time model output by increasing the order of the GQ approximation. When we explored higher orders in our simulation study, we found very little benefit to orders greater than five.

For a discrete-time step Markov model, it is important to have a correction method which can be easily applied. Our GQ method can be applied as a post-hoc correction to an uncorrected Markov trace, and can be easily implemented in spreadsheet software such as Microsoft Excel. In previous comparisons of existing methods, Simpson's 1/3 rule gave the best approximation to the continuous-time model, [5, 18] and our results support this finding.

In our simple numerical example, the standard half-cycle correction yielded a *worse* approximation than the uncorrected output from the discrete-time model. This is a concern, and suggests that half-cycle correction method should be avoided.

7.2 How this fits with the existing literature

There are a number of papers in the literature on correction methods for discrete time models, see, for example [5, 6, 18–20]. The main published methods are the half-cycle correction, the trapezoidal and Simpson's methods, and the life-table method. We did not include the life-table method in our analysis because it has been shown to be equivalent to the trapezoidal method (and therefore also to the half-cycle method). [5, 18]

Elbasha and Chhatwal (2016) highlighted the need to test a range of other methods, including Gaussian quadrature, Simpson's 6-points rule, Boole's rule and Romberg's quadrature rule. However, in all of the literature so far, the target for approximation has been the cumulative total outcome (costs and QALYs in our example). What is unique about our approach is that we use Gaussian quadrature to approximate a *correction factor* (that post-multiplies the Markov trace) rather than approximating the cumulative total outcome itself. Our simulation study suggests that even a low-order (i.e. 3rd-order) GQ approximation results in an excellent correction.

7.3 Strengths and limitations

The GQ method is easy to implement in both spreadsheet software and in R, and can be applied as a *post hoc* correction to uncorrected Markov traces. We provide an Excel spreadsheet and R implementations as supplementary materials. Extension to a higher orders than the 5th is straightforward, though the benefit is likely to be inconsequential.

Our study has limitations. We have not computed bounds for the relative errors, or established rigorous criteria for when the 5th order GQ is guaranteed to outperform other methods. Furthermore, although we undertook a PSA to investigate the robustness of our results to different parameters, we used only a single model. We did not investigate the effect of different state-space dimension on our approximation's performance. The model type considered for the simulation study was one with the constant transition matrix rather than time-varying transition matrix. While it is not obvious how to generalize our approximation scheme directly to the time-varying case, use of tunnel states can approximate time dependence by increasing the dimension of the model state-space.

7.4 Problems with discrete-time models

Every rate matrix $\mathbf{R}$ gives rise to a unique transition probability matrix $\mathbf{M}$ for time step t via $\mathbf{M} = \exp(\mathbf{R}t)$. This means that every continuous-time model can be expressed as a discrete-time model. However, the converse is not necessarily true. Given $\mathbf{M}$, there may not be a unique rate matrix $\mathbf{R}$, or $\mathbf{R}$ may not exist at all (see e.g. Higham [21]). Indeed, determining whether a transition matrix $\mathbf{M}$ has a unique rate matrix $\mathbf{R}$, is an unsolved problem in probability theory, where it is known as the 'embedding' (or 'embeddability') problem. [22]

This presents a specific problem that arises in practice, namely determining the transition matrix for shorter cycle length than originally specified (eg. the monthly transition matrix for a model defined by a yearly transition matrix). The typical approach to this is to compute the p-th root of the matrix, where p is the multiplicative factor between the time periods (e.g. 12 for converting the yearly cycle model to the monthly cycle model). Even when it is defined (which is not guaranteed), the p-th root of $\mathbf{M}$ may not be a valid transition matrix. In the context of financial applications, methods have been developed that choose approximate p-th roots for models specified via transition matrices, [23, 24] however models specified in this way remain unsatisfactory.

All these issues are absent for models that are defined in terms of a rate matrix $\mathbf{R}$; transition matrices over any time period are guaranteed to be valid, and are unique. This is an important motivation for modelling healthcare decision problems in terms of their continuous-time dynamics, rather than artificially as discrete-time systems, but the historical reliance on spreadsheet applications has rather hampered the adoption of continuous time models.

We would argue that defining models in continuous time is a more natural and logically consistent representation of reality, and obviates the need for correction methods or stochastic simulation. Often data on the efficacy of interventions are reported as hazard ratios, which can be applied more naturally to continuous-time models defined in terms of rates. Methods for estimating rates from discrete-time data are available, [9, 25] including as an R package. [26]

Finally, where cycle corrections are employed, we would caution against their unthinking use. Some costs may be periodic and aligned with cycle length, eg. injections with a set frequency. These costs are incurred by all relevant compliant patients at the start of a cycle; if cycle correction are performed, intervention costs will be underestimated.

7.5 Conclusion

A novel Gaussian quadrature-based method for correcting the output of a discrete-time Markov model provides a better approximation to an assumed underlying continuous-time Markov model than do current methods.

Authors' Note : This work was done in ScHARR, University of Sheffield, UK. An abstract from this work was presented as a poster presentation at the International Society of Pharmacoeconomics and Outcomes Research (ISPOR) - EU, Denmark in November 2019

Conflict of Interest and Financial Support: The author(s) declared no conflicts of interest with respect to the research, authorship, and/or publication of this article. This is an independent research. TS was supported by the National Institute for Health Research (NIHR) Fellowship (NIHR300461). The views expressed in this publication are those of the author and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health and Social Care. PJD was supported by a fellowship from the UK Medical Research Council (MR/P022081/1); this UK funded award is part of the EDCTP2 programme supported by the European Union.

References

[1] Uwe Siebert, Oguzhan Alagoz, Ahmed M Bayoumi, Beate Jahn, Douglas K Owens, David J Cohen, Karen M Kuntz, and ISPOR-SMDM Modeling Good Research Practices Task Force. State-transition modeling: A report of the ISPOR-SMDM Modeling Good Research Practices Task Force–3. *Value Health*, 15(6):812–820, 2012.

[2] A Briggs and M Sculpher. An introduction to Markov modelling for economic evaluation. *PharmacoEconomics*, 13(4):397–409, 1998.

[3] F A Sonnenberg and J R Beck. Markov models in medical decision making: a practical guide. *Med Decis Making*, 13(4):322–338, 1993.

[4] David M J Naimark, Michelle Bott, and Murray Krahn. The half-cycle correction explained: Two alternative pedagogical approaches. *Med Decis Making*, 28(5):706–712, 2008.

[5] Elamin H Elbasha and Jagpreet Chhatwal. Theoretical Foundations and Practical Applications of Within-Cycle Correction Methods. *Med Decis Making*, 36(1):115–131, 2016.

[6] Jan J Barendregt. The half-cycle correction: Banish rather than explain it. *Med Decis Making*, 29(4):500–502, 2009.

[7] Dirk P Laurie. Computation of Gauss-type quadrature formulas. *J Comput Appl Math*, 127(1):201–217, 2001.

[8] E B Dynkin. Kolmogorov and the Theory of Markov Processes. *Ann Probab*, 17(3):822–832, 1989.

[9] Nicky J Welton and A E Ades. Estimation of markov chain transition probabilities and rates from fully and partially observed data: uncertainty propagation, evidence synthesis, and model calibration. *Med Decis Making*, 25(6):633–645, 2005.

[10] Edmund Jones, David Epstein, and Leticia García-Mochón. A Procedure for Deriving Formulas to Convert Transition Rates to Probabilities for Multistate Markov Models. *Med Decis Making*, 37(7):779–789, 2017.

[11] H E Moore. On the reciprocal of the general algebraic matrix. *Bull Am Math Soc.*, 26:394–395, 1920.

[12] Kendall E Atkinson. *An Introduction to Numerical Analysis*. John Wiley & Sons, 2008.

[13] Daniel J Velleman. The Generalized Simpson's Rule. *Am Math Mon.*, 112(4):342–350, 2005.

[14] Philip J Davis and Philip Rabinowitz. *Methods of Numerical Integration*. Academic Press, 1984.

[15] Adi Ben-Israel and Thomas N E Greville. *Generalized Inverses: Theory and Applications*. Springer Science & Business Media, 2003.

[16] Arthur Wouk. Integral representation of the logarithm of matrices and operators. *Journal of Mathematical Analysis and Applications*, 11:131 – 138, 1965.

[17] Andrew H Briggs, Milton C Weinstein, Elisabeth A L Fenwick, Jonathan Karnon, Mark J Sculpher, A David Paltiel, and ISPOR-SMDM Modeling Good Research Practices Task Force. Model parameter estimation and uncertainty: A report of the ISPOR-SMDM Modeling Good Research Practices Task Force–6. *Value Health*, 15(6):835–842, 2012.

[18] Elamin H Elbasha and Jagpreet Chhatwal. Myths and misconceptions of Within-Cycle correction: A guide for modelers and decision makers. *PharmacoEconomics*, 34(1):13–22, 2016.

[19] David M J Naimark, Nader N Kabboul, and Murray D Krahn. The half-cycle correction revisited: Redemption of a kludge. *Med Decis Making*, 33(7):961–970, 2013.

[20] David M J Naimark, Nader N Kabboul, and Murray D Krahn. Response to "The life table method of half-cycle correction: getting it right". *Med Decis Making*, 34(3):286–287, 2014.

[21] Nicholas J. Higham. *Functions of matrices : theory and computation*. Society for Industrial and Applied Mathematics, Philadelphia, 2008.

[22] Chen Jia. A solution to the reversible embedding problem for finite markov chains. *Statistics & Probability Letters*, 116:122 – 130, 2016.

[23] Robert B Israel, Jeffrey S Rosenthal, and Jason Z Wei. Finding Generators for Markov Chains via Empirical Transition Matrices, with Applications to Credit Ratings. *Math Finance*, 11(2):245–265, 2001.

[24] Alexander Kreinin and Marina Sidelnikova. Regularization Algorithms for Transition Matrices. *Algo Research Quarterly*, 4(1):23–40, 2001.

[25] Yasunari Inamura. Estimating Continuous Time Transition Matrices From Discretely Observed Data. Technical Report 06-E-7, Bank of Japan, 2006.

[26] Marius Pfeuffer. `ctmcd`: An R Package for Estimating the Parameters of a Continuous-Time Markov Chain from Discrete-Time data. *The R Journal*, 9(2):127, 2017.

A Appendix: Justification for the use of the Moore-Penrose generalised inverse

We wish to establish a sufficient condition for

$$\mathbf{s}^\top \left(\sum_{t=0}^{N-1} \mathbf{M}^t \right) \mathbf{c} = \mathbf{s}^\top (\mathbf{M}^N - \mathbb{I})(\mathbf{M} - \mathbb{I})^+ \mathbf{c}, \tag{30}$$

to be true.

As in the paper, we will write $\mathbf{A} = (\mathbf{M} - \mathbb{I})$, and $\mathbf{A}^+$ for the Moore-Penrose inverse of $\mathbf{A}$. We recognise that Equation 30 is valid if:

$$\left(\sum_{t=0}^{N-1} \mathbf{M}^t - (\mathbf{M}^N - \mathbb{I})\mathbf{A}^+ \right) \mathbf{c} = \mathbf{0}.$$

We can show this is true if we can show that $\mathbf{c} = \mathbf{A}\mathbf{A}^+\mathbf{c}$. This is because, if $\mathbf{c} = \mathbf{A}\mathbf{A}^+\mathbf{c}$, we can then write

$$\begin{aligned} &\left(\sum_{t=0}^{N-1} \mathbf{M}^t - (\mathbf{M}^N - \mathbb{I})\mathbf{A}^+ \right) \mathbf{c} \\ &= \left(\sum_{t=0}^{N-1} \mathbf{M}^t - (\mathbf{M}^N - \mathbb{I})\mathbf{A}^+ \right) \mathbf{A}\mathbf{A}^+\mathbf{c} \\ &= \left\{ (\mathbb{I} + \mathbf{M} + \cdots + \mathbf{M}^{N-1})\mathbf{A} - (\mathbf{M}^N - \mathbb{I}) \right\} \mathbf{A}^+\mathbf{c} \\ &= \left\{ (\mathbb{I} + \mathbf{M} + \cdots + \mathbf{M}^{N-1})(\mathbf{M} - \mathbb{I}) - (\mathbf{M}^N - \mathbb{I}) \right\} \mathbf{A}^+\mathbf{c} \\ &= \left\{ (\mathbf{M}^N - \mathbb{I}) - (\mathbf{M}^N - \mathbb{I}) \right\} \mathbf{A}^+\mathbf{c} \\ &= \mathbf{0}. \end{aligned}$$

Note that $\mathbf{A}\mathbf{A}^+$ is a projection matrix because it is a property of the Moore-Penrose generalised inverse that $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$, and therefore that $(\mathbf{A}\mathbf{A}^+)^2 = \mathbf{A}\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}\mathbf{A}^+$. Moreover, $\mathbf{A}\mathbf{A}^+$ is a projector onto the column space of $\mathbf{A}$, because $\mathbf{A}\mathbf{A}^+\mathbf{A} = \mathbf{A}$. So our desired Equation 30 holds if $\mathbf{c}$ is in the column space of $\mathbf{A}$.

A state in a Markov chain is defined as *recurrent* if the probability of returning to this state at some future iteration is 1; otherwise it is *transient*. The set S of states for the Markov chain decomposes into the set of transient states, T, and recurrent states, C, ie $S = T \cup C$. It is known that the set C is closed, meaning that once a Markov chain enters C, it does not leave it. In health economics, the most common type of recurrent state is an absorbing state (a single state that is never left), often representing 'dead'. It is natural to assume that costs are zero for recurrent states so that sums of costs for cohorts over time converge; otherwise, one would be assuming some form of cost-accruing immortality. As well as being a natural assumption, we now show that zero costs for recurrent states ensures that $\mathbf{c}$ is in the column space of $\mathbf{A}$, and therefore allows us to use the Moore-Penrose inverse.

First, note that we can reorder states to bring any Markov chain transition matrix into the canonical form (i.e. all recurrent states first, and then afterwards any transient states). We can write this matrix in block form as

$$\mathbf{M} = \left[\begin{array}{c|c} \mathbf{M}_{CC} & \mathbf{0} \\ \hline \mathbf{M}_{TC} & \mathbf{M}_{TT} \end{array} \right], \tag{31}$$

which makes clear how starting points in T eventually end up in C. No movement from C to T is possible because the probabilities associated with these transitions (which appear in the upper right block) are all zero.

While $\mathbf{M}_{CC}$ must itself be a stochastic matrix (i.e. the rows sum to 1), $\mathbf{M}_{TT}$ is not; the sums along its rows are *less* than 1, reflecting the 'leakage' from T to C. We can write $\mathbf{M}_{TT} = \mathbf{D}\tilde{\mathbf{M}}_{TT}$

with $\mathbf{D}$ diagonal and $\mathbf{D}_{ii} = \sum_{j \in T} p_{ij} < 1$ the probability of remaining in T from state i in the next iteration, and the stochastic matrix $\tilde{\mathbf{M}}_{TT}$ defined by dividing the i-th row of $\mathbf{M}_{TT}$ by $\mathbf{D}_{ii}$. Since the maximum absolute eigenvalue of a stochastic matrix is 1 and the eigenvalues of $\mathbf{D}$ are < 1, the matrix $\mathbf{M}_{TT}$ has absolute eigenvalues < 1.

Note that $\mathbf{A}$ has the same block structure as $\mathbf{M}$ in Equation 31, i.e.

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{A}_{CC} & \mathbf{0} \\ \hline \mathbf{A}_{TC} & \mathbf{A}_{TT} \end{array}\right]. \tag{32}$$

Because $\mathbf{M}_{CC}$ is a stochastic matrix, $\mathbf{A}_{CC} = (\mathbf{M}_{CC} - \mathbb{I})$ has rows that sum to zero. This means that it must have at least one column that is a linear combination of the remaining columns, and is therefore singular (i.e. not invertible). In contrast, because the eigenvalues of $\mathbf{M}_{TT}$ are < 1, $\mathbf{A}_{TT} = (\mathbf{M}_{TT} - \mathbb{I})$ is non-singular and therefore full rank. This implies that for any $\mathbf{c}$ that is zero on C, i.e.,

$$\mathbf{c} = \left[\begin{array}{c} \mathbf{0} \\ \hline \mathbf{c}_T \end{array}\right],$$

then $\mathbf{c}$ is in the column space of $\mathbf{A}$ since $\mathbf{0}$ is in the column space of all matrices, and $\mathbf{c}_T$ is in the column space of $\mathbf{A}_{TT}$. To see why $\mathbf{c}_T$ is in the column space of $\mathbf{A}_{TT}$, recall that the column space of an $n \times n$ non-singular matrix is $\mathbb{R}^n$, and therefore includes all n-dimensional vectors.

Thus, we have shown that if $\mathbf{c}$ is zero for recurrent states, this is a sufficient condition to allow us to use Equation 30.

B Appendix: Proof that Simpson's 1/3 rule and the second-order Gaussian quadrature are equivalent

Our second order approximation for costs is

$$T_c^{GQ(2)} = \sum_{t=0}^{N-1} \mathbf{s}_0^\top \mathbf{M}^t \hat{\mathbf{Z}}_2^{-1} \mathbf{c},$$

where $\hat{\mathbf{Z}}_2$ is given by

$$\hat{\mathbf{Z}}_2 = \frac{1}{2}\left(\frac{1+\sqrt{\frac{1}{3}}}{2}\mathbf{M} + \frac{1-\sqrt{\frac{1}{3}}}{2}\mathbb{I}\right)^{-1} + \frac{1}{2}\left(\frac{1-\sqrt{\frac{1}{3}}}{2}\mathbf{M} + \frac{1+\sqrt{\frac{1}{3}}}{2}\mathbb{I}\right)^{-1}.$$

Using standard manipulations, one can show that

$$\hat{\mathbf{Z}}_2^{-1} = \frac{1}{3}\left(\mathbf{M}^2 + 4\mathbf{M} + \mathbb{I}\right)(\mathbf{M} + \mathbb{I})^{-1}.$$

Note we can also write this as

$$\hat{\mathbf{Z}}_2^{-1} = \frac{1}{2}[(\mathbb{I} + \mathbf{M})^2 - (1/3)(\mathbb{I} - \mathbf{M})^2](\mathbb{I} + \mathbf{M})^{-1}$$

$$= \frac{1}{2}[(\mathbb{I} + \mathbf{M}) - (1/3)(\mathbb{I} - \mathbf{M})(\mathbb{I} + \mathbf{M})^{-1}(\mathbb{I} - \mathbf{M})],$$

since $(\mathbb{I} + \mathbf{M})^{-1}$ commutes with $(\mathbb{I} - \mathbf{M})$.

Defining

$$\sigma_\alpha = \sum_{t=0}^{N-1} \alpha^t \mathbf{M}^t,$$

note that

$$\sigma_{+1}(\mathbb{I}-\mathbf{M}) = \sigma_{-1}(\mathbb{I}+\mathbf{M}) = \mathbb{I}-\mathbf{M}^N, \tag{33}$$

with the last equality holding for even N, and also that

$$T_c^{GQ(2)} = \mathbf{s}_0^\top \left(\sigma_{+1}\mathbf{F}\right)\mathbf{c}, \tag{34}$$

where we can expand the parenthesis in Equation 34 using Equation 33 to find

$$\begin{aligned}
\sigma_{+1}\mathbf{F} &= \frac{1}{2}\sigma_{+1}(\mathbb{I}+\mathbf{M}) - \frac{1}{6}(\mathbb{I}-\mathbf{M}^N)(\mathbb{I}+\mathbf{M})^{-1}(\mathbb{I}-\mathbf{M}) \\
&= \frac{1}{2}\sigma_{+1}(\mathbb{I}+\mathbf{M}) - \frac{1}{6}\sigma_{-1}(\mathbb{I}-\mathbf{M}) \\
&= \frac{1}{2}\sum_{t=0}^{N-1}\left(1-\frac{(-)^t}{3}\right)\mathbf{M}^t + \frac{1}{2}\sum_{t=0}^{N-1}\left(1+\frac{(-)^t}{3}\right)\mathbf{M}^{t+1} \\
&= \frac{1}{3}\mathbf{M}^0 + \frac{2}{3}\sum_{\substack{t>0\\ \text{even}}}^{N-2}\mathbf{M}^t + \frac{4}{3}\sum_{\substack{t=1\\ \text{odd}}}^{N-1}\mathbf{M}^t + \frac{1}{3}\mathbf{M}^N,
\end{aligned} \tag{35}$$

since

$$1+\frac{(-)^{t+1}}{3} = 1-\frac{(-)^t}{3} = \begin{cases} 2/3, & \text{if } t \text{ even} \\ 4/3, & \text{if } t \text{ odd} \end{cases}.$$

Substituting Equation 35 back for the parenthesis in Equation 34 proves the claimed equivalence of GQ2 and Simpson's 1/3 rule.

Supplemental Document

Improving Cycle Corrections in Discrete Time Markov Models: A Gaussian Quadrature Approach

Tushar Srivastava, Mark Strong, Matthew D Stevenson, Peter J Dodd

S1 Impact of time horizon on model error

As a scenario analysis we have varied time horizon of the model and analyse the impact on absolute error between continuous-time and discrete-time model adjusted with different cycle correction methods. Figure S1 presents the analysis result. It can be seen that the higher order GQ method outperformed other existing methods at every instance.

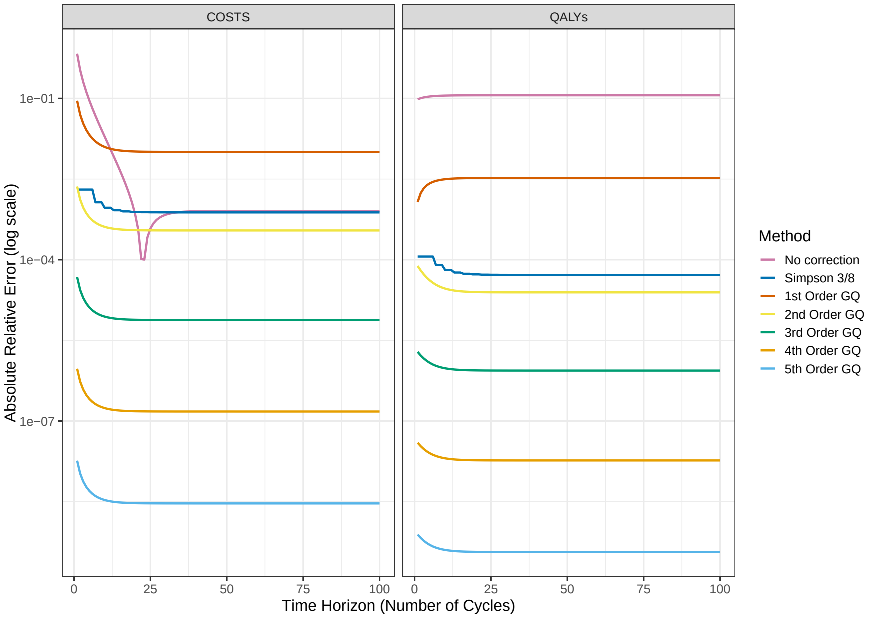

Figure S1: Impact of time horizon on absolute error between continuous- and discrete-time model using different approximation methods. 1st order GQ is same as HCC and trapezoidal and 2nd order GQ is same as Simpson 1/3. Clearly higher order GQ method has lowest error compared to the other existing methods. Simpson 3/8 method was run only for time horizons which were multiple of 3. Abbreviations: GQ: Gaussian Quadrature; HCC: Half-Cycle Correction; QALY: Quality Adjusted Life-Years.

S2 Discounted results

Table S1 shows the discounted results. 3.5 % annual discount rate was considered for both costs and QALYs.

Correction Method	Costs	Absolute Error	QALYs	Absolute Error	NMB
Continuous time model	£190.5293	-	3.73378	-	-
5th order GQ method	£190.5293	8.48e-09	3.73378	1.26e-07	-£0.1341
4th order GQ method	£190.5293	2.74e-07	3.73378	1.57e-07	-£0.1375
3rd order GQ method	£190.5270	1.20e-05	3.73377	1.39e-06	-£0.2739
2nd order GQ method	£190.4363	4.88e-04	3.73368	2.62e-05	-£2.9653
Simpson's 3/8 method	£190.3289	1.05e-03	3.73357	5.40e-05	-£5.9629
1st order GQ method	£188.2323	1.21e-02	3.75069	4.53e-03	£509.5421
No correction performed	£190.7323	1.07e-03	4.22569	1.32e-01	£14,757.0421

Table S1: Comparative analysis discounted result per person based on different methods. Absolute error is presented in increasing order of error. Discounting was performed before cycle correction. 5th order GQ method has the absolute minimum error for both costs and QALYs and closest to continuous model. 1st order GQ is same as HCC and trapezoidal and 2nd order GQ is same as Simpson 1/3. NMB is with respect to continuous time model with willingness-to-pay of £30,000. Abbreviations: GQ: Gaussian Quadrature; HCC: Half-Cycle Correction; NMB: Net Monetary Benefit; QALY: Quality Adjusted Life-Years

S3 Where is 5th order GQ less likely to be the most accurate approximation?

Figure S2 shows which parameter values are less likely to result in 5th order GQ being the best approximation method. Figure S3 explores interactions effects, ie. whether certain pairwise combinations of parameters are less likely to result in 5th order GQ being the best approximation method.

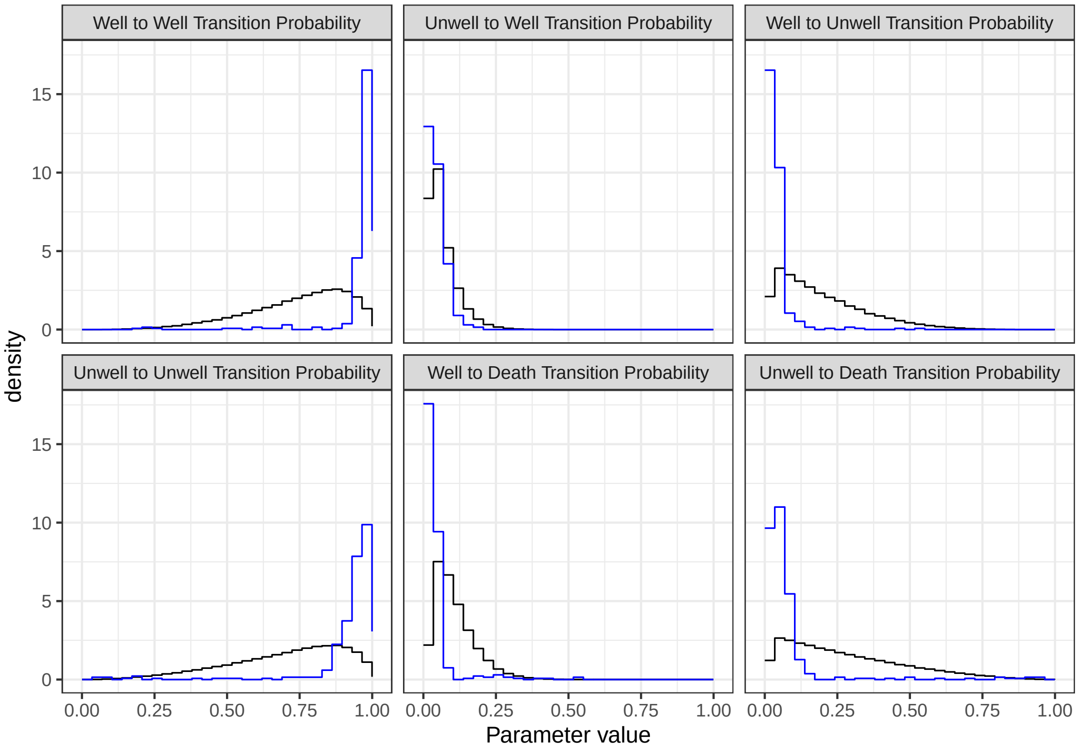

Figure S2: Parameter histograms when 5th order GQ is not best in 100,000 simulations (blue), compared to sampled parameters (black).

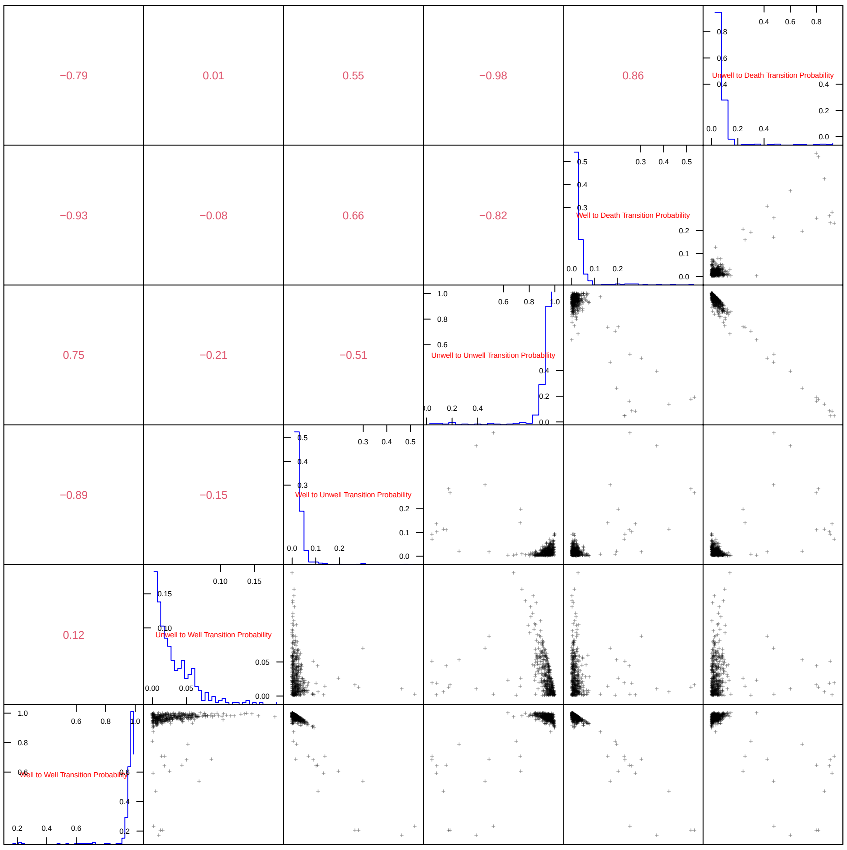

Figure S3: Pairs plot of parameters for which 5th order GQ is not best in 100,000 simulations.

S4 Expressions for $\hat{\mathbf{Z}}_n$

Given an unadjusted discrete-time Markov model

$$T_c^{DT} = \sum_{t=0}^{N-1} \mathbf{s}_0^{'} \mathbf{M}^t \mathbf{c},$$

the n-th order Gaussian quadrature approximation to the continuous-time model output is given by

$$T_c^{GQ(n)} = \sum_{t=0}^{N-1} \mathbf{s}_0^{'} \mathbf{M}^t \hat{\mathbf{Z}}_n^{-1} \mathbf{c},$$

where $\hat{\mathbf{Z}}_n$ is defined as

$$\hat{\mathbf{Z}}_n = \sum_{i=1}^{n} w_i f\left(\frac{1+x_i}{2}\right).$$

The function $f(\cdot)$ is defined as

$$\begin{aligned} f(x) &= \{(\mathbf{M} - \mathbb{I})x + \mathbb{I}\}^{-1} \\ &= \{x\mathbf{M} + (1-x)\mathbb{I}\}^{-1}. \end{aligned}$$

Locations $\{x_i\}$ and weights $\{w_i\}$ for the first 5 orders of Gaussian-Legendre quadrature are

Order $\{n\}$	Locations $\{x_i\}$	Weights $\{w_i\}$
1	0	2
2	$\pm\sqrt{\frac{1}{3}}$	1
3	0	$\frac{8}{9}$
	$\pm\sqrt{\frac{3}{5}}$	$\frac{5}{9}$
4	$\pm\sqrt{\frac{3}{7}-\frac{2}{7}\sqrt{\frac{6}{5}}}$	$\frac{18+\sqrt{30}}{36}$
	$\pm\sqrt{\frac{3}{7}+\frac{2}{7}\sqrt{\frac{6}{5}}}$	$\frac{18-\sqrt{30}}{36}$
5	0	$\frac{128}{225}$
	$\pm\frac{1}{3}\sqrt{5-2\sqrt{\frac{10}{7}}}$	$\frac{322+13\sqrt{70}}{900}$
	$\pm\frac{1}{3}\sqrt{5+2\sqrt{\frac{10}{7}}}$	$\frac{322-13\sqrt{70}}{900}$

This leads the following expressions for $\hat{\mathbf{Z}}_n$ for $n = 1, \ldots, 5$:

$$\hat{\mathbf{Z}}_1 = 2(\mathbf{M} + \mathbb{I})^{-1}.$$

$$\begin{aligned}
\hat{\mathbf{Z}}_2 &= \frac{1}{2}\left\{ f\left(\frac{1+\sqrt{\frac{1}{3}}}{2}\right) + f\left(\frac{1-\sqrt{\frac{1}{3}}}{2}\right)\right\} \\
&= \frac{1}{2}\left(\frac{1+\sqrt{\frac{1}{3}}}{2}\mathbf{M} + \frac{1-\sqrt{\frac{1}{3}}}{2}\mathbb{I}\right)^{-1} + \frac{1}{2}\left(\frac{1-\sqrt{\frac{1}{3}}}{2}\mathbf{M} + \frac{1+\sqrt{\frac{1}{3}}}{2}\mathbb{I}\right)^{-1},
\end{aligned}$$

$$\begin{aligned}
\hat{\mathbf{Z}}_3 &= \frac{1}{2}\left\{ \frac{8}{9} f\left(\frac{1+0}{2}\right) + \frac{5}{9} f\left(\frac{1+\sqrt{\frac{3}{5}}}{2}\right) + \frac{5}{9} f\left(\frac{1-\sqrt{\frac{3}{5}}}{2}\right)\right\} \\
&= \frac{8}{9}(\mathbf{M} + \mathbb{I})^{-1} + \frac{5}{18}\left(\frac{1+\sqrt{\frac{3}{5}}}{2}\mathbf{M} + \frac{1-\sqrt{\frac{3}{5}}}{2}\mathbb{I}\right)^{-1} + \frac{5}{18}\left(\frac{1-\sqrt{\frac{3}{5}}}{2}\mathbf{M} + \frac{1+\sqrt{\frac{3}{5}}}{2}\mathbb{I}\right)^{-1},
\end{aligned}$$

$$\begin{aligned}
\hat{\mathbf{Z}}_4 &= \frac{1}{2}\left\{ \frac{18+\sqrt{30}}{36} f\left(\frac{1+\sqrt{\frac{3}{7}-\frac{2}{7}\sqrt{\frac{6}{5}}}}{2}\right) + \frac{18+\sqrt{30}}{36} f\left(\frac{1-\sqrt{\frac{3}{7}-\frac{2}{7}\sqrt{\frac{6}{5}}}}{2}\right)\right\} \\
&\quad + \frac{1}{2}\left\{ \frac{18-\sqrt{30}}{36} f\left(\frac{1+\sqrt{\frac{3}{7}+\frac{2}{7}\sqrt{\frac{6}{5}}}}{2}\right) + \frac{18-\sqrt{30}}{36} f\left(\frac{1-\sqrt{\frac{3}{7}+\frac{2}{7}\sqrt{\frac{6}{5}}}}{2}\right)\right\} \\
&= \frac{1}{2}\left\{ \frac{18+\sqrt{30}}{36}\left(\frac{1+\sqrt{\frac{3}{7}-\frac{2}{7}\sqrt{\frac{6}{5}}}}{2}\mathbf{M} + \frac{1-\sqrt{\frac{3}{7}-\frac{2}{7}\sqrt{\frac{6}{5}}}}{2}\mathbb{I}\right)^{-1}\right\} \\
&\quad + \frac{1}{2}\left\{ \frac{18+\sqrt{30}}{36}\left(\frac{1-\sqrt{\frac{3}{7}-\frac{2}{7}\sqrt{\frac{6}{5}}}}{2}\mathbf{M} + \frac{1+\sqrt{\frac{3}{7}-\frac{2}{7}\sqrt{\frac{6}{5}}}}{2}\mathbb{I}\right)^{-1}\right\} \\
&\quad + \frac{1}{2}\left\{ \frac{18-\sqrt{30}}{36}\left(\frac{1+\sqrt{\frac{3}{7}+\frac{2}{7}\sqrt{\frac{6}{5}}}}{2}\mathbf{M} + \frac{1-\sqrt{\frac{3}{7}+\frac{2}{7}\sqrt{\frac{6}{5}}}}{2}\mathbb{I}\right)^{-1}\right\} \\
&\quad + \frac{1}{2}\left\{ \frac{18-\sqrt{30}}{36}\left(\frac{1-\sqrt{\frac{3}{7}+\frac{2}{7}\sqrt{\frac{6}{5}}}}{2}\mathbf{M} + \frac{1+\sqrt{\frac{3}{7}+\frac{2}{7}\sqrt{\frac{6}{5}}}}{2}\mathbb{I}\right)^{-1}\right\},
\end{aligned}$$

$$
\begin{aligned}
\hat{\mathbf{Z}}_5 &= \frac{1}{2}\left\{\frac{128}{225} f\left(\frac{1+0}{2}\right)\right\} \\
&\quad + \frac{1}{2}\left\{\frac{322+13\sqrt{70}}{900} f\left(\frac{1+\frac{1}{3}\sqrt{5-2\sqrt{\frac{10}{7}}}}{2}\right) + \frac{322+13\sqrt{70}}{900} f\left(\frac{1-\frac{1}{3}\sqrt{5-2\sqrt{\frac{10}{7}}}}{2}\right)\right\} \\
&\quad + \frac{1}{2}\left\{\frac{322-13\sqrt{70}}{900} f\left(\frac{1+\frac{1}{3}\sqrt{5+2\sqrt{\frac{10}{7}}}}{2}\right) + \frac{322-13\sqrt{70}}{900} f\left(\frac{1-\frac{1}{3}\sqrt{5+2\sqrt{\frac{10}{7}}}}{2}\right)\right\} \\
&= \frac{128}{225}(\mathbf{M}+\mathbb{I})^{-1} \\
&\quad + \frac{1}{2}\left\{\frac{322+13\sqrt{70}}{900}\left(\frac{1+\frac{1}{3}\sqrt{5-2\sqrt{\frac{10}{7}}}}{2}\mathbf{M} + \frac{1-\frac{1}{3}\sqrt{5-2\sqrt{\frac{10}{7}}}}{2}\mathbb{I}\right)^{-1}\right\} \\
&\quad + \frac{1}{2}\left\{\frac{322+13\sqrt{70}}{900}\left(\frac{1-\frac{1}{3}\sqrt{5-2\sqrt{\frac{10}{7}}}}{2}\mathbf{M} + \frac{1+\frac{1}{3}\sqrt{5-2\sqrt{\frac{10}{7}}}}{2}\mathbb{I}\right)^{-1}\right\} \\
&\quad + \frac{1}{2}\left\{\frac{322-13\sqrt{70}}{900}\left(\frac{1+\frac{1}{3}\sqrt{5+2\sqrt{\frac{10}{7}}}}{2}\mathbf{M} + \frac{1-\frac{1}{3}\sqrt{5+2\sqrt{\frac{10}{7}}}}{2}\mathbb{I}\right)^{-1}\right\} \\
&\quad + \frac{1}{2}\left\{\frac{322-13\sqrt{70}}{900}\left(\frac{1-\frac{1}{3}\sqrt{5+2\sqrt{\frac{10}{7}}}}{2}\mathbf{M} + \frac{1+\frac{1}{3}\sqrt{5+2\sqrt{\frac{10}{7}}}}{2}\mathbb{I}\right)^{-1}\right\}.
\end{aligned}
$$

S5 Distribution of input parameters

Figure S4 presents distribution of input parameters of the model considered for probabilistic sensitivity analysis.

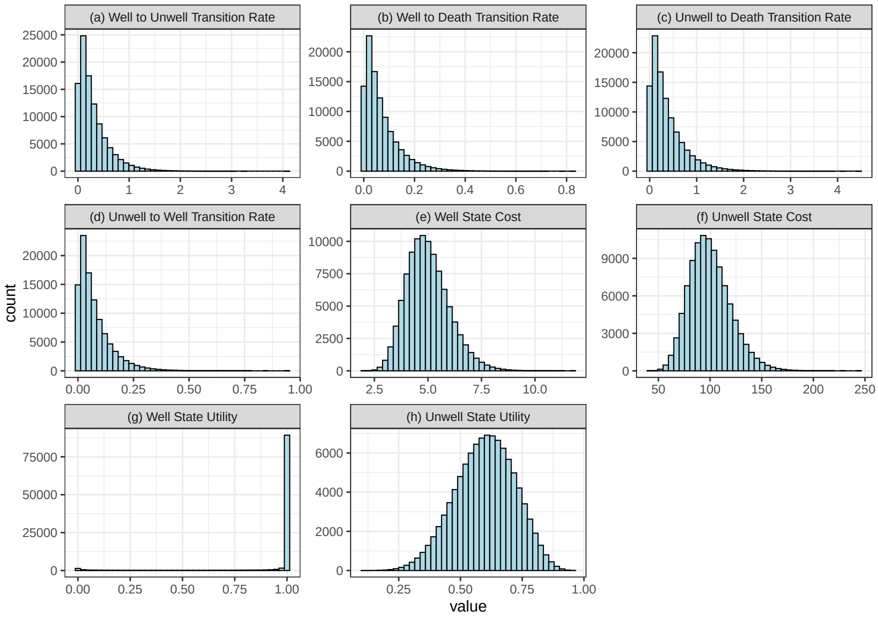

Figure S4: Distribution used in PSA for each model input parameters (100,000 samples were used).Abbreviation: PSA: Probabilistic Sensitivity Analysis