
Article:
Srivastava, T. orcid.org/0000-0002-5961-9348, Latimer, N.R. orcid.org/0000-0001-5304-5585 and Tappenden, P. orcid.org/0000-0001-6612-2332 (2021) Estimation of transition probabilities for state-transition models : a review of NICE appraisals. PharmacoEconomics, 39 (8). pp. 869-878. ISSN 1170-7690

https://doi.org/10.1007/s40273-021-01034-5

Article Type: Practical Application

Manuscript Title: Estimation of Transition Probabilities for State-Transition Models: A Review of NICE Appraisals

Running Title (45 characters)**:** Transition Probability in State Transition Models

Authors:

Tushar Srivastava, MSc ; School of Health and Related Research (ScHARR), University of Sheffield, UK (Orcid ID: 0000-0002-5961-9348)

Nicholas R Latimer, PhD ; School of Health and Related Research (ScHARR), University of Sheffield, UK (Orcid ID: 0000-0001-5304-5585)

Paul Tappenden, PhD ; School of Health and Related Research (ScHARR), University of Sheffield, UK (Orcid ID: 0000-0001-6612-2332)

Corresponding Authors Details:

Tushar Srivastava,

School of Health and Related Research (ScHARR),

University of Sheffield, UK, S1 4DA

t.srivastava@sheffield.ac.uk

Declarations

Funding:

TS was supported by the National Institute for Health Research (NIHR) Fellowship (NIHR300461). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health and Social Care. NRL is supported by Yorkshire Cancer Research (Award S406NL).

Conflicts of interest/Competing interests:

NL reports personal fees from Pierre Fabre and Merck Sharp & Dohme outside the submitted work.

Author Contributions:

All authors (TS, NRL, PT) contributed to the study conception and design. Data extraction, and analysis were performed by TS in close discussion with NRL and PT. The first draft of the manuscript was written by TS and all authors commented on and edited previous versions of the manuscript. All authors read and approved the final manuscript.

Availability of data and material:

All data analysed during this review are included in this published article [and its supplementary information files] and are available in public domain in NICE website.

Code availability:

Not applicable

Words Count: 4582

Number of Pages: 15

Number of Figures: 4

Number of Tables: 2

Appendices/Supplemental Materials: Tables - 3 ; Pages - 12

Abstract

State transition models (STM) are used to inform health technology reimbursement decisions. Within STMs, the movement of patients between the model health states over discrete time intervals is determined by transition probabilities (TPs). Estimating TPs presents numerous issues, including missing data for specific transitions, data incongruence and uncertainty around extrapolation. Inappropriately estimated TPs could result in biased models. There is limited guidance on how to address common issues associated with TP estimation. In order to assess current methods for estimating TPs and to identify issues that may introduce bias, we reviewed National Institute for Health and Care Excellence (NICE) Technology Appraisals (TAs) published from 1st January 2019 to 27th May 2020. Twenty-eight models (from 26 TAs) were included in the review. Several methods for estimating TPs were identified: survival analysis (n=11); count method (n=9); multi-state modelling (n=7); logistic regression (n=2); negative binomial regression (n=2); Poisson regression (n=1); and calibration (n= 1). Evidence Review Groups identified several issues relating to TP estimation within these models, including important transitions being excluded (n=5); potential selection bias when estimating TPs for post-randomisation health states (n=2); issues concerning the use of multiple data sources (n=4); potential biases resulting from the use of data from different populations (n =2), and inappropriate assumptions around extrapolation (n =3). These issues remained unresolved in almost every instance. Failing to address these issues may bias model results and lead to sub-optimal decision-making. Further research is recommended to address these methodological problems.

Key Points for the Decision Makers:

- State transition models (STMs) are the most common type of health economic models used in medical decision-making.
- Patients' movement between the model health states over discrete time intervals is determined by transition probabilities (TP).
- Our review of NICE Technology Appraisals demonstrates that while some potentially appropriate methods have been used for estimating TPs in STMs, there still exist various issues in TP estimation.

1. Introduction

1.1 Background

Health economic models are commonly required to assess the cost-effectiveness of alternative competing health care technologies.[1] The need to use models largely stems from the limitations of randomised controlled trials (RCT) in providing all necessary information to measure and value all relevant costs and health outcomes for all treatment options. Health economic modelling can play a crucial role in structuring the decision problem, extrapolating short-terms outcomes beyond the observed period of a clinical trial, synthesising evidence of relative treatment effects for all comparators across multiple studies and allowing for the full characterisation of decision uncertainty.[1,2]

The state transition model (STM) is the most common type of model used in medical decision-making because of its simplicity in describing complex real-life phenomena.[3,4,5] STMs define an underlying disease process in terms of a series of mutually exclusive and jointly exhaustive disease – and/or treatment-related states (e.g., well, unwell, dead) and provide a basis for estimating different trajectories through these states for patients receiving different health care interventions. The movement of patients between the model health states over discrete time intervals is determined by transition probabilities (TPs). TPs may be constant with respect to time, or may be assumed to be conditional on time – either on time since entry into the model (e.g., varying with the age of the patient cohort) or on time since entry into a particular health state (e.g., event risks vary according to time elapsed since a prior clinical event occurred). STMs can be evaluated using cohort simulation in discrete time (reflecting mean event risks across a population) or using a patient-level simulation approach. The simplest form of STM is one in which the matrix of transition rates, and hence TPs, is fixed in every time interval, giving rise to a Markov chain which is time-homogeneous (time invariant). This approach may be extended to reflect time-varying TPs through Markov processes and semi-Markov approaches in which event risks are conditional on time since model entry or on time since entry into an intermediate health state.[1,6]

TPs included in STMs are often derived from analyses of individual patient data (IPD), which can come from RCTs or non-randomised studies such as observational studies or registries. Summary data from these sources are also regularly used to estimate TPs and, in some instances, published literature and expert opinion may also be used. Given that TPs represent the means of estimating patient trajectories through the model health states, they represent a crucial component of the model itself. If TPs are estimated inappropriately, this may introduce bias into the model, which may, in turn, lead to erroneous cost-effectiveness results and sub-optimal adoption decisions.

TPs can be estimated using a variety of parametric and non-parametric methods.[1,7,8] For example, survival modelling may be used as a parametric approach, while the count method may be used as a non-parametric approach.[1] However, there are several limitations in applying these methods that arise either from restrictive assumptions underpinning the method or limitations regarding the data used to estimate the TPs themselves. As such, model developers may turn to methodological guidelines to understand best practice in TP estimation. However, current guidance is focused on health economic model development, and rigorous methodological guidelines on how to estimate TPs are lacking.[9] A recent review by Olariu et al. (2017) indicated no consensus statement or guideline on how to estimate TPs.[10]

Alternative techniques for estimating TPs may have an important impact on cost-effectiveness estimates. For example, in National Institute for Health and Care Excellence (NICE) Technology Appraisal (TA) Number 443 (obeticholic acid

for primary biliary cholangitis), there was disagreement between the Evidence Review Group (ERG) and the company concerning the most appropriate method for TP estimation.[11] The ERG preferred to estimate TPs directly from data in the pivotal RCT, whereas the company's submitted model used calibration methods to estimate TPs indirectly. In an exploratory analysis undertaken by the company and the ERG, the incremental cost-effectiveness ratio (ICER) was subject to considerable variation, ranging from 8% to 141% compared to the base case in different scenarios, when TPs were estimated using data directly from pivotal RCT approach. There are numerous other examples where there is uncertainty around TP estimation.

The purpose of this paper is to identify current approaches for estimating TPs in STMs and to identify related problems and challenges. To achieve this aim, we undertook a review of recent NICE TAs that were completed during the period 2019 to 2020. We specifically examined which data sources were used to inform TPs, the types of analytical methods used to estimate TPs and issues identified by independent ERGs and NICE Appraisal Committees (ACs) relating to these methods. Error and uncertainty may be introduced at various points in the model development process, including during model structuring, implementation and parameterisation. In this paper, we focus on errors introduced when estimating TPs. This is not because other sources of error and uncertainty are less important, but rather because sensitivity analysis is commonly used to assess the importance of structural and parameter uncertainty, whereas errors or uncertainty associated with the methods used to estimate TPs are not routinely taken into account. This review is intended to provide a basis for directing future research on TPs' estimation in STMs.

1.2. Potential issues with estimating TPs from data

To describe the potential problems associated with TP estimation, we use a hypothetical example of an STM (Figure 1). In this example model, four health states are included: 'No Symptoms', 'Mild/Moderate Disease', 'Severe Disease' and 'Dead', and only forward transitions are possible (i.e., patients cannot revert to better health states). Despite this simplicity, there can be multiple issues in estimating TPs in this scenario.

Issue 1 - Missing transitions. Suppose we have observed data on some of the transitions; however, some transitions are missing either because patients at certain severity levels were not included (or followed up) in the trial or because certain transitions are rare. For example, there may be available data on $P_{1,n}$ and $P_{2,n}$ but trial patients were no longer followed up (or censored) when they developed severe disease, hence $P_{3,n}$ is missing. Alternatively, the study may have recruited patients with mild/moderate disease only; hence, information to inform all TPs from the 'No Symptoms' health state is missing. In such instances, we may need to synthesise evidence from multiple sources to estimate the trace across all states included in the STM. This overlaps with another issue concerning the use of multiple data sources for TP estimation.

Issue 2 – Sources of data. Information on the inputs for TP estimation for all transitions ($P_{m,n}$) is frequently obtained from various data sources with varying follow-up durations, baseline characteristics, and study types (e.g. RCTs, non-randomized observational studies, registries etc.). However, it is challenging to combine these data statistically or use them directly for populating transitions. For example, data from a non-randomised observational study may be available for informing transition $P_{2,3}$. However, the patients' baseline characteristics in this study (e.g. median age 30 years) may differ from those in the pivotal RCT used to inform other transitions (e.g. median age 65 years), making its adjustment

difficult for TP estimation. Another problem arises when several estimates are available for a certain transition, for example, $P_{1,2}$, as this may lead to challenges in selecting or synthesising a preferred estimate to include in the model.

Issue 3 - Data missingness. Suppose that data are available for all TPs ($P_{1,n}$ to $P_{3,n}$), but some individuals' observations are missing. For example, some patients with severe disease may be too sick to attend assessments; hence $P_{3,3}$ and $P_{3,4}$ (observed TPs) are subject to informative censoring, and available estimates may not be generalisable across the whole target population. In such cases, assumptions are required regarding data missingness to avoid TP estimation bias, and using alternative assumptions leads to different TP estimates.

Issue 4 - Data on subgroups unavailable. Suppose we have access to data to inform all TPs ($P_{1,n}$ to $P_{3,n)}$ for the overall target population, but not for all individual subgroups. If we assume that TPs are equivalent between the subgroups and the overall population, this may lead to bias. Conversely, using data from some (available) subgroups may lead to other issues such as missing transitions and small sample sizes.

Issue 5 - Need for extrapolation. Suppose that we have access to TPs ($P_{1,n}$ to $P_{3,n)}$ over the duration of the pivotal trial, but the trial stopped before all patients experienced the events of interest. In this case, we need to extrapolate TPs beyond the observed period of the trial, including making assumptions about the relative treatment effects between competing decision options. The choices regarding the relationship between event risk and time will impact on modelled health state occupancy at later timepoints.

Issue 6 - Long intervals between assessments. To inform TPs in STMs, mismatches between the data collection interval and the model interval are common. For example, in our observed data, patients may be followed up every six months. However, if we use a six-monthly model cycle, this may lead to a lack of precision in the model estimates; hence, we may need to adopt a shorter model cycle length.

Issue 7 - Data incongruence. Suppose we have published data reporting on time to a particular event (e.g., time to death $P_{1,4}$, $P_{2,4}$ and $P_{3,4}$), but the model needs to capture the underlying disease process (i.e., all TPs $P_{1,n}$ to $P_{4,n}$). To estimate the underlying transition rates, we would need to use the information we have available (or observed) to estimate the information we do not have available (or cannot observe), given the model structure e.g., through model calibration.

<FIGURE 1>

2. Methods

This review focussed on STMs submitted as part of the NICE TA programme. We considered NICE for its rigour in producing guidance on new and existing health technologies (such as drugs and medical devices) and its focus on economic value assessments of health technologies. NICE guidance reflects the outcome of important decisions about the use of health technologies in real-world scenarios. Also, submissions to NICE are freely accessible online. Most submissions to NICE use cutting-edge methods and NICE regularly updates its methods guidelines which are followed widely across the world.

2.1 Study eligibility criteria

We undertook a review of TAs for which guidance documents were published from 1st January 2019 to 27th May 2020. This review period was selected pragmatically to identify a sufficient number of appraisals that would provide a representative sample of the TP estimation methods currently used in health economic models. Company submissions, ERG reports and Final Appraisal Determination (FAD) documents from the NICE website were reviewed. We extracted information on issues raised by the ERGs or ACs from these documents. Appraisal documents were included in the review if they presented cost-effectiveness analyses using STMs. The full inclusion and exclusion criteria for the review are presented in Table 1.

2.2 Data extraction and presentation of results

A data extraction form was designed in Microsoft Excel (Microsoft Corporation, Redmond, WA, USA). Information was extracted on parameters such as: data sources used to inform TPs for models; methods used for TP estimation; issues with TP estimation raised by the ERG and/or the ACs, and general details (e.g., disease area, model type, number of model health states). The extraction form was piloted in two TAs; this was subsequently reviewed and further refined by the study authors.

We provided a brief overview of each identified method used for TP estimation in a summary table to clarify on terminology. We extracted data on TP estimation issues, as described by the ERG and/or ACs as free text. We categorised the data sources and methods used for TP estimation in included models and presented these using doughnut charts.

2.3 Quality assessment

We did not formally assess the quality of the included models. This is because the purpose of the review was to identify TP estimation issues, rather than to scrutinise the quality of ERG reports or NICE FADs.

<TABLE 1>

3. Results

3.1. Summary of included studies

We screened a total of 78 TAs published within the considered timeframe (January 2019 to May 2020). Of these, 28 models (within 26 TAs) met the inclusion criteria (Figure 2). Terminated TAs (n=14) were excluded from the review. The majority of the remaining excluded models were partitioned survival models (26 out of 38 TAs), which do not use TPs to estimate health state occupancy. Eleven included models were in cancer; the remaining 17 models were in other disease areas, including multiple sclerosis and Crohn's disease. There were no dynamic models among the models included - all were static. In terms of the nature of TPs considered, almost all included models involved time-varying TPs either according to model time or time since entry into an intermediate health state (using a semi-Markov approach).

A full list of the reviewed models and a summary of each model is presented in supplementary materials (Suppl. Table 1, Suppl. Table 2).

<FIGURE 2>

3.2. Data and methods used to estimate transition probabilities

3.2.1 Data sources

Across most of the models (n=28), multiple sources were required to populate the full set of TPs included in the models (Figure 3). In four models (TA605, TA593, TA587, TA578), TPs were estimated solely from the pivotal RCT. Most models used IPD from the pivotal RCT supplemented with data from non-randomised observational studies (either IPD or aggregate data) to estimate TPs. All-cause mortality data from national life tables were commonly used as a constraint for modelling mortality or to model mortality risk in people with non-terminal diseases (e.g., TA613, TA614) in the identified models.

<FIGURE 3>

3.2.2 Methods used to estimate TPs

Several methods were used to estimate TPs in the included models (Figure 4). The most common approaches were the non-parametric count method (in non-cancer models) (n=9) and parametric survival modelling methods (in cancer models) (n=11). Other approaches used were: multi-state modelling (MSM) (n=7); multinomial logistic regression (n=2); negative binomial regression (n=2); Poisson regression (n=1) and calibration (n=1). In one model (TA607), the specific method applied was unclear from the ERG report and the FAD. A summary of each method is provided in Table 2.

<FIGURE 4>
<TABLE 2>

3.3 Issues identified with transition probability estimation

Across the 28 included models, ERGs identified several problems related to the estimation of TPs. In all cases, the NICE AC agreed with the ERG's concerns on TP-related issues. Overall, these issues can be divided into three of the categories previously identified: (i) sources of data, (ii) missing transitions, and (iii) extrapolation.

3.3.1 Sources of data:

The most common issues were related to the use of inappropriate data sources to inform TPs. The first issue is the risk of mismatching populations, due to the use of multiple or different sources of data to inform the TPs. For example, in TA623 (patiromer for hyperkalaemia), real-world data (RWD) were considered for informing some transitions. The ERG noted that patients' baseline characteristics in the pivotal RCT were different from those in the RWD. However, due to the lack of IPD available from the RWD, the ERG could not assess the impact of this issue on the ICER. In another example, TA580 (enzalutamide for prostate cancer), the company used data from another published trial of enzalutamide instead of the pivotal enzalutamide RCT to estimate some TPs. In the ERG's exploratory analysis, the ICER increased from £28,853 to £31,671 per QALY gained (above NICE's usual cost-effectiveness threshold) when the pivotal enzalutamide RCT was used as the source for TP estimation. TA569 (pertuzumab for HER2-positive early-

stage breast cancer) and TA624 (peginterferon beta-1a for multiple sclerosis) are further examples where ERGs noted mismatching populations between TP data sources.

The second data source related issue concerns bias that arises in TP estimates from RCT data. In typical three-state cancer models, defined in terms of progression-free (PF), progressed disease (PD) and dead, TP estimates for the transition from PD to dead may subject to post-randomisation bias.[12] Post-randomisation bias can occur when data from an RCT are analysed using a starting point that is after the randomisation time-point. For example, if cancer RCT randomised patients when they are progression-free, estimates of TPs from the PF health state benefit from randomisation and therefore represent the full trial population and provide unbiased relative TPs between treatment groups. In contrast, if not all patients have experienced disease progression at the time of analysis, TPs estimated for the PD health state will not be representative of the overall trial population, and may result in biased relative treatment effects if different proportions of patients have experienced disease progression in each treatment group. Specifically, if not all patients experience disease progression during the RCT, and if rapid progressors have a different survival prognosis to those who progress later, fitting parametric survival models to post-progression survival (PPS) data may introduce selection bias and informative censoring, potentially leading to erroneous clinical benefits and cost-effectiveness estimates. [13, 14] The ERGs explicitly raised this issue in TA578 (durvalumab for non-small cell lung cancer [NSCLC]) and TA593 (ribociclib, for breast cancer); however, this issue has broader relevance to most cancer STMs when estimates are based on censored RCT data. ERGs cited NICE Technical Support Document (TSD) 19 for details on this issue.[13] However, as no method has yet been developed to adjust for PPS bias, ERGs did not attempt to perform any exploratory analysis to understand its impact on the ICERs.

The third issue concerns bias that arises in TP estimates from non-randomised observational study data. To estimate TPs using non-randomised observational data, ERGs noted that adjustments such as matching patient baseline characteristics between the observational data and other sources of data used to estimate TPs in the model (e.g., the pivotal RCT) were usually not made, and therefore were not completely consistent with the statistical adjustment methods recommended for this purpose in NICE TSD 17, resulting in the possibility of selection bias and confounding (TA558, TA604, TA614 and TA616).[15]

The last issue concerns small sample sizes, as ERGs raised concerns around the precision of estimated TPs. In TA565 (ribociclib with fulvestrant for breast cancer), the ERG noted that the TPs used in the company's analysis might not be robust given the relatively small sample sizes (n = 47 to 136) used to obtain these estimates. Furthermore, the ERG could not verify the company's treatment effectiveness analysis, as IPD were not provided. Another example is TA578 (durvalumab for NSCLC), in which the ERG raised concerns about the small number of progressed patients through which to estimate PPS. PPS data were immature, and there was substantial uncertainty around its extrapolation. To address some of the uncertainty in PPS, the ERG explored alternative assumptions by selecting an alternative parametric survival model to represent PPS and found that the ICER varied from £49,868 to £59,131 per QALY gained (the base case ICER was £50,238 per QALY gained). Small sample size concerns were also noted in TA614 (cannabidiol for Dravet syndrome), where patients in the study trial were mostly under 18 years of age (98.11%). The model population was stratified into four age groups (2-5, 6-11, 12-17, and 18-55 years). However, using the study population estimate

for the 18-55 years age group raised concern around biased TPs because this group represented only 1.89% of the study population.

3.3.2 Missing Transitions:

Missing transitions, i.e., ignoring relevant transitions that are likely to occur in reality, but were not included in the company's model, represented another common issue reported by ERGs. A frequently reported reason for excluding these transitions was that they were not observed in the dataset used to inform the TPs (usually the pivotal RCT). However, if these transitions are possible but have not been observed, then applying a zero probability will bias the model results. This issue was raised as a concern in TA556, TA590, TA605, TA607, and TA623. In exploratory analyses conducted by ERGs, it was found that assigning non-zero TPs to the missing transitions led to considerable variation in the ICER, ranging from -24% to 202% relative to the company's base case.

3.3.3. Extrapolation

Extrapolation was one of the most significant concerns raised by ERGs, as this requires strong assumptions about TPs beyond the observed period of RCTs or other studies used to inform the model. The ERGs highlighted that some of these assumptions might be inappropriate. For example, in TA588 (nusinersen for spinal muscular atrophy), the ERG was concerned that the company's assumptions about the intervention's continued effectiveness beyond the observed period of the relevant RCTs were highly optimistic. The ERG explored these assumptions and found that the ICER increased by a factor between 1.2 and 40 compared with the company's base case estimate. In other appraisals, when parametric survival models informed the extrapolation, ERGs found that the ICER was very sensitive to the parametric model choice. For instance, in TA556 (darvadstrocel for Crohn's disease), the use of different parametric survival models for the remission and relapse time-to-event functions led to ICERs ranging from £20,591 to £133,311 per QALY gained. This issue was also raised by the NICE AC, who were concerned about the uncertainty around the long-term benefit and cost-effectiveness of the treatment.

Assumptions regarding the duration of treatment effects beyond observed trial follow-up were also key factors impacting the extrapolation. In an exploratory analysis by the ERG in TA569 (pertuzumab for breast cancer), the ICER increased by 60% when a different assumption on treatment effect waning was considered compared to the company's base case. Similarly, in TA578, the ERG explored assumptions around treatment effect waning in the extrapolated period and found that the ICER varied from £47,000 to £64,531 per QALY gained, with ERG preferred base case of £50,238 per QALY gained. With no treatment waning effect, the ICER was £60,928 per QALY gained. Another concern for ERGs was when constant TPs were assumed during extrapolation for some transitions, as reported in TA556 and TA569.

4. Discussion

Our review of NICE TAs demonstrates that various data sources and methods are used to estimate TPs for STMs. IPD obtained from RCTs and non-randomised observational studies were the most commonly used data sources for TP estimation. Survival analysis and non-parametric count methods were the most commonly used approaches for estimating TPs. In the included models, ERGs identified several important issues with the TP estimation. The key issues

were related to data sources used for TP estimation, missing transitions and extrapolations. These issues ultimately impact model results and decision-making.

Despite the widespread use of STMs in health economic evaluation, guidance on how to estimate TPs is limited. A recent review by Olariu et al (2017) found no consensus statements or guidelines regarding TP estimation in STMs.[9] The review identified one relevant publication in which limited guidance was provided on the use of rates and probabilities. Also, a recent tutorial published by Gidwani et al. (2020) provided recommendations on TP estimation when data to derive TPs are in the form of: relative risks, odds, odds ratios or rates.[16] These limited recommendations are helpful, but several challenges in TP estimation remain.

Other researchers have made a significant contribution in providing solutions to some of the issues in TP estimation. For example, Williams et al. (2016), Putter et al. (2007) contributed to methods based on MSMs.[17,18] However, MSMs are not commonly used in health economics. Chhatwal et al (2016) and Craig et al (2002) provided method of estimation when TPs are from different sources with varying follow-up durations or intervals (Suppl. Table 3).[19,20] However, the generalisability and applicability of these solutions are still unexplored, and their recommendations have not been implemented yet.

We recommend issues around TP estimation related to multiple data sources, extrapolation, and post-randomised health states should be considered as priorities for further methodological research. Analysts will frequently encounter multiple data sources with mismatching populations when estimating TPs. The inappropriate use of data for estimating TPs can impact on modelled on cost-effectiveness outcomes (as observed in this review) and ultimately, decision making. Extrapolation is a key element of cost-effectiveness assessments. Guidance regarding survival model selection for extrapolation is available for individual and independently modelled survival endpoints;[21] however, guidance to inform survival model selection in STM settings is scarce. Further work in this area would be valuable. In addition, it is very common for STMs to include TPs from health states that patients only enter after a period of time – for example, once disease has progressed. Careful thought is required regarding the appropriateness of the data and methods used to estimate TPs from these health states. If data from RCTs are used, the potential presence and impact of selection bias and informative censoring should be considered. Research on methods to address these issues is needed. Similarly, if observational data are used to inform TPs from late-occurring health states, potential population mismatches must be considered, as should the relevance of the data source(s) to the target population under consideration. Finally, it is important to recognise that TPs estimated from RCT datasets represent transitions observed in trial sample populations, which may not fully represent the target population for whom treatment recommendations will be made. Therefore, the use of Bayesian methods for estimating TPs representing disease populations should be explored.

To the best of our knowledge, this is the first attempt to identify the frequency of TP estimation related issues in NICE appraisals. Given the lack of recommendations and guidance in this area, it is important to properly identify problems associated with TP estimation, to highlight this for analysts and decision-makers, and to provide a basis for subsequent research. Although we have focused on TPs in this paper, some of the issues raised (e.g., sources of data, data missingness) are also relevant for other model parameters such as costs and utilities; hence, the implications are broad.

Our study is subject to some limitations. To make our review concise and pragmatic, we only considered recent TAs, i.e., from January 2019 to May 2020. Consequently, our review found examples of only three of the seven potential issues (sources of data, missing transitions, and extrapolation) associated with estimating TPs that we identified prior to conducting our review. The other four issues (data missingness, data on subgroup unavailable, long intervals between assessments, and data incongruence) were not identified in the review; however, we believe these remain important concerns for TP estimation. We may have missed some other techniques used to estimate TPs, or additional issues that ERGs may have raised in earlier appraisals. In addition, TAs were reviewed and checked by only one reviewer (TS); however, substantive discussions and debate took place between co-authors during the review process. Notably, we have only reported on TP estimation issues that ERGs and ACs identified. There is a possibility that issues may have been identified which were not included in published reports and documents or that other issues existed that were not identified by ERGs or ACs. However, the scope of our review did not include an assessment of whether ERGs and/or ACs correctly identified or addressed specific issues; rather, we sought to provide a complete summary of the issues associated with TP estimation reported in TAs.

TP estimation is very common in other areas, such as medical statistics, finance, agriculture, computer science and engineering. We suggest that expertise from these broader areas can be borrowed in health economics to enhance TP estimation and hence, healthcare decision-making.

5. Conclusion

Problems associated with TP estimation are common, as observed in NICE TAs. It is important to address these issues to provide unbiased TP estimates. Failing to address these issues may result in biased model results, leading to sub-optimal decisions. Further research is required to address these methodological problems.

6. References

[1] Briggs A; Sculpher M, Claxton K. Decision Modelling for Health Economic Evaluation, Oxford University Press; 2006

[2] Buxton MJ, Drummond MF, Van Hout BA, et al. Modelling in economic evaluation: an unavoidable fact of life. *Health Econ*. 1997;6(3):217-227. doi:10.1002/(sici)1099-1050(199705)6:3<217::aid-hec267>3.0.co;2-w

[3] Briggs A, Sculpher M. An introduction to Markov modelling for economic evaluation. *Pharmacoeconomics*. 1998;13(4):397-409. doi:10.2165/00019053-199813040-00003

[4] Siebert U, Alagoz O, Bayoumi AM, et al. State-transition modeling: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--3. *Value Health*. 2012;15(6):812-820. doi:10.1016/j.jval.2012.06.014

[5] Sonnenberg FA, Beck JR. Markov models in medical decision making: a practical guide. *Med Decis Making*. 1993;13(4):322-338. doi:10.1177/0272989X9301300409

[6] Hawkins N, Sculpher M, Epstein D. Cost-effectiveness analysis of treatments for chronic disease: using R to incorporate time dependency of treatment response. Med Decis Making. 2005;25(5):511-519. doi:10.1177/0272989X05280562

[7] Welton NJ, Ades AE. Estimation of markov chain transition probabilities and rates from fully and partially observed data: uncertainty propagation, evidence synthesis, and model calibration. *Med Decis Making*. 2005;25(6):633-645. doi:10.1177/0272989X05282637

[8] 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):625-641. doi:10.1177/0272989X10384738

[9] Caro JJ, Briggs AH, Siebert U, Kuntz KM; ISPOR-SMDM Modeling Good Research Practices Task Force. Modeling good research practices--overview: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-1. Med Decis Making. 2012;32(5):667-677. doi:10.1177/0272989X12454577

[10] Olariu E, Cadwell KK, Hancock E, Trueman D, Chevrou-Severac H. Current recommendations on the estimation of transition probabilities in Markov cohort models for use in health care decision-making: a targeted literature review. *Clinicoecon Outcomes Res*. 2017;9:537-546. Published 2017 Sep 1. doi:10.2147/CEOR.S135445

[11] NICE Single Technology Appraisal. Obeticholic acid for treating primary biliary cholangitis (TA 443). NICE. 2017. [Available from: https://www.nice.org.uk/guidance/ta443] Accessed 25 July 2020

[12] Fergusson D, Aaron SD, Guyatt G, Hébert P. Post-randomisation exclusions: the intention to treat principle and excluding patients from analysis. BMJ. 2002;325(7365):652-654. doi:10.1136/bmj.325.7365.652

[13] Woods B, Sideris E, Palmer S, Latimer N, Soares M. NICE DSU Technical Support Document 19. Partitioned Survival Analysis for Decision Modelling in Health Care: A Critical Review. 2017 [Available from http://www.nicedsu.org.uk] Accessed 15 July 2020

[14] García-Albéniz X, Maurel J, Hernán MA. Why post-progression survival and post-relapse survival are not appropriate measures of efficacy in cancer randomized clinical trials. Int J Cancer. 2015;136(10):2444-2447. doi:10.1002/ijc.29278

[15] Faria, R., Hernandez Alava, M., Manca, A., Wailoo, A.J. NICE DSU Technical Support Document 17: The use of observational data to inform estimates of treatment effectiveness for Technology Appraisal: Methods for comparative individual patient data. 2015.[Available from http://www.nicedsu.org.uk] Accessed 15 July 2020

[16] Gidwani R, Russell LB. Estimating Transition Probabilities from Published Evidence: A Tutorial for Decision Modelers [published correction appears in Pharmacoeconomics. 2020 Sep 8]. Pharmacoeconomics. 2020;38(11):1153-1164. doi:10.1007/s40273-020-00937-z

[17] Williams C, Lewsey JD, Briggs AH, Mackay DF. Cost-effectiveness Analysis in R Using a Multi-state Modeling Survival Analysis Framework: A Tutorial. Med Decis Making. 2017;37(4):340-352. doi:10.1177/0272989X16651869

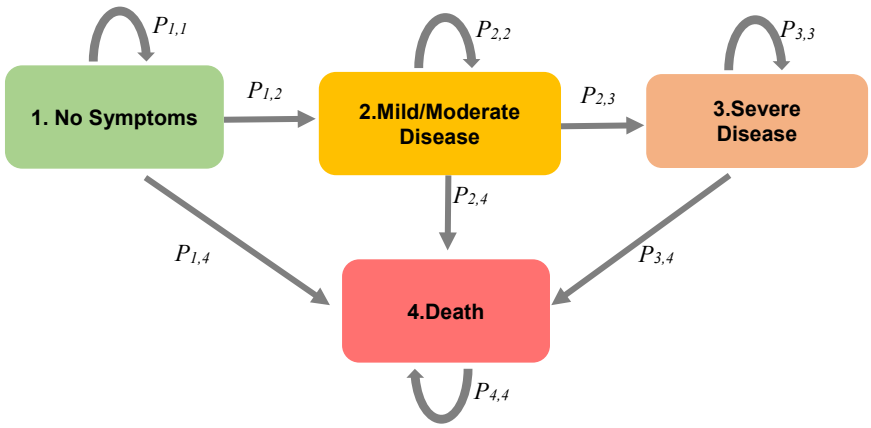
[18] Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Stat Med*. 2007;26(11):2389-2430. doi:10.1002/sim.2712

[19] Chhatwal J, Jayasuriya S, Elbasha EH. Changing Cycle Lengths in State-Transition Models: Challenges and Solutions. *Med Decis Making*. 2016;36(8):952-964. doi:10.1177/0272989X16656165

[20] Craig BA, Sendi PP. Estimation of the transition matrix of a discrete-time Markov chain. *Health Econ*. 2002;11(1):33-42. doi:10.1002/hec.654

[21] Latimer, N. NICE DSU Technical Support Document 14: Undertaking survival analysis for economic evaluations alongside clinical trials - extrapolation with patient-level data. 2011. [Available from http://www.nicedsu.org.uk] Accessed 15 July 2020

Figure 1 : Example state transition model structure and permitted transition probabilities



Note: Prs is a transition probability, which describes the likelihood of moving from state m to state n ≠ m at time t over a model cycle of length *c*, and $P_{m,m}(t)$ describes the likelihood of remaining in state m at time t over a model cycle of length *c*.

Figure 2: Flow diagram of NICE Technology Appraisal inclusion and exclusion

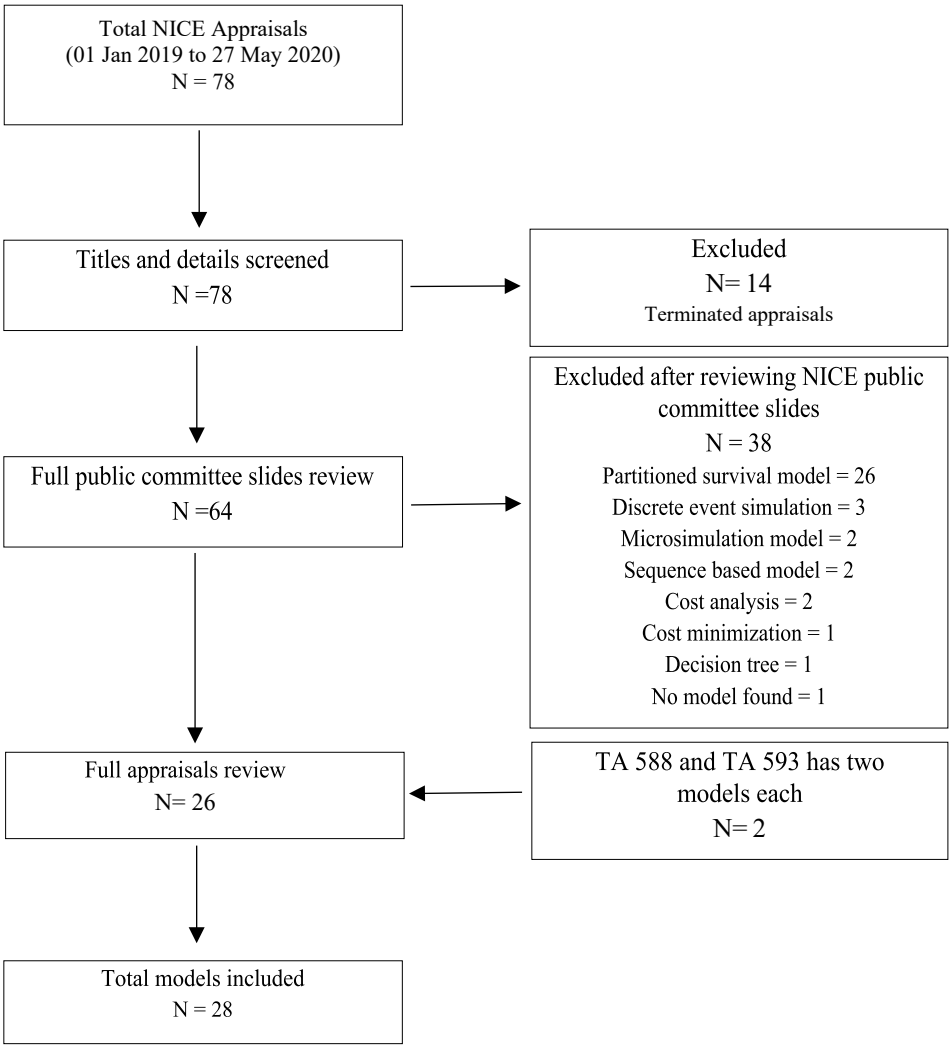
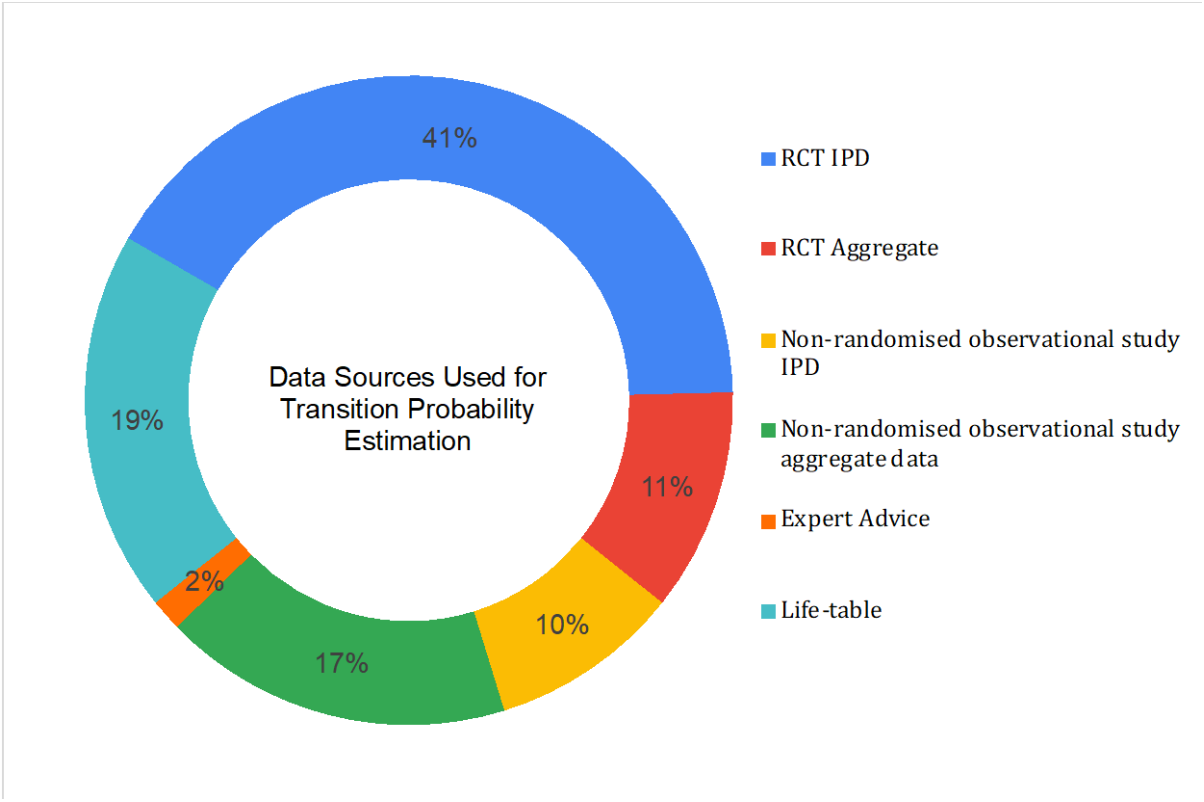


Figure 3 : Data Sources used in NICE Technology Appraisals for informing transition probability estimation



Abbreviations: IPD: Individual Patient Data; RCT: Randomised controlled trial

Figure 4: Methods used in NICE Technology Appraisals for transition probability estimation

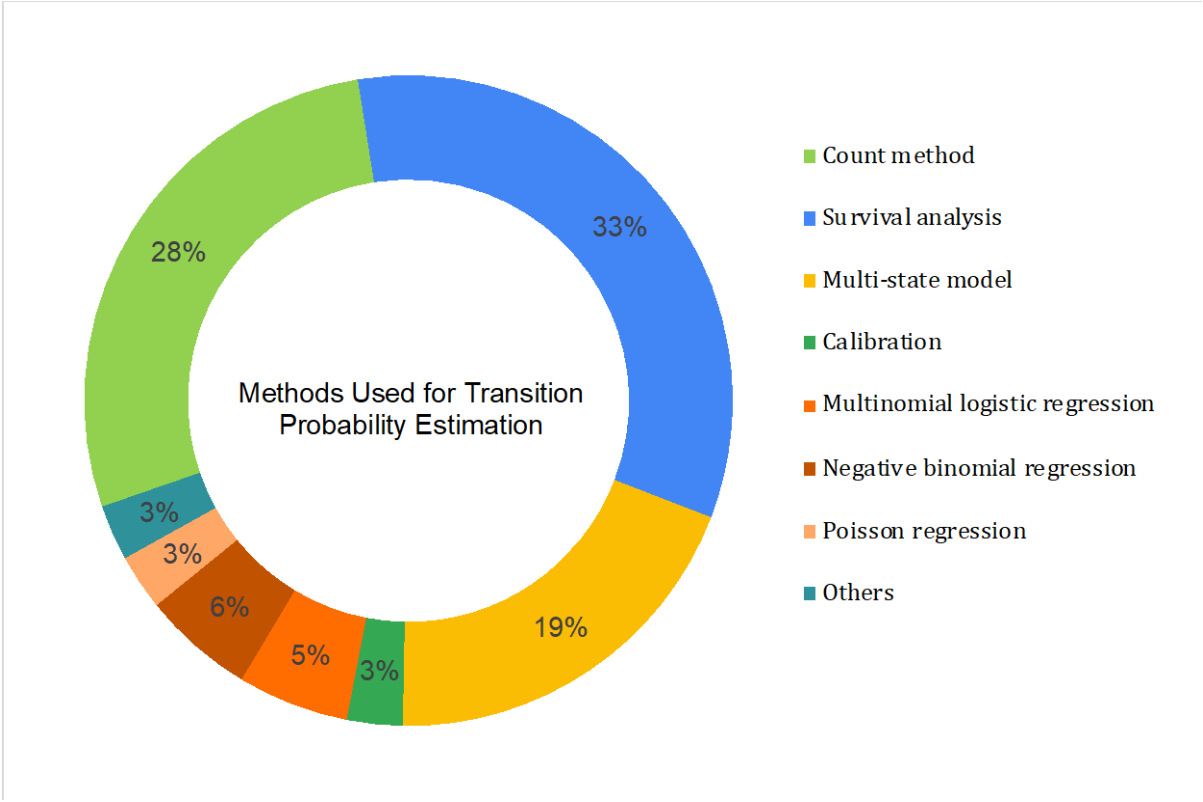


Table 1: Inclusion and exclusion criteria for the review

Category	Inclusion Criteria	Exclusion Criteria
Economic evaluation type	Cost-effectiveness analysis; Cost-utility analysis	Cost-analysis; Cost-minimisation analysis; Cost-consequence analysis; Budget impact analysis
Modelling type	State transition models (including Markov chain, Markov process, semi-Markov and multistate models)	Partitioned survival model; Microsimulation model; Discrete event simulation model; Decision tree; Sequence-based model with no transitions
Issues related to transition probabilities	Evidence Review Group (ERG) reports; NICE Final Appraisal Documents (FADs)	Terminated Technology Appraisals

Table 2 : Summary of methods used to estimate transition probabilities

Methods	Description
Non-parametric count method	This method is relatively straightforward if the states' sequence for each individual observation is observed, i.e., if the individual transitions are observed. The method requires patients to have pairs of consecutive observations. The probability of transition from any given state i is equal to the proportion of individuals that started in state i and ended in state j as a proportion of all individuals that started in state i.
Survival modelling	Parametric survival functions are fitted to IPD typically using maximum likelihood estimation (MLE) to estimate the survival function parameters (e.g., the scale and shape of Weibull distribution) which is then used to derive TPs. The baseline TP of the event of interest is defined as one minus the ratio of the survivor function at the end of the interval to the survivor function at the beginning of the interval ($TP_{baseline} = 1 - \frac{S(ti)}{S(ti-1)}$; (t_{i-1}, t_i) is time interval)
Multi-state modelling	A multi-state model framework describes how an individual move between a series of states in continuous time. This approach models each of the transitions of interest simultaneously and could be considered when there are a series of competing events and when these events occur sequentially. R packages such that *mstate* and *msm* can be used to implement the multi-state model to estimate TPs. The *msm* function fits the model to the available time-to-event data directly using maximum likelihood estimation. The TPs are estimated endogenously within the *msm* function. In contrast, the *mstate* package estimates TPs exogenously from the survival function and combines them under a competing risk framework.
Logistic regression	The logistic regression model is used to define covariate-dependent TPs in the presence of only two possible discrete outcomes. It is an appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Generally, in health economic models with multiple health states, multinomial logistic regression model used for TP estimation. Multinomial logistic regression is a classification method that generalizes logistic regression to multiclass problems, i.e., with more than two possible discrete outcomes. The model is used to estimate the effect of independent variables and predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables.
Negative binomial regression	In the presence of a small sample and the need to account for over- or under-dispersion by modelling the variance independently of the mean of the data, negative binomial regression is a good choice for TP estimation. It is a type of generalized linear model in which the dependent variable is a count of the number of times an event occurs.
Model calibration	Calibration is useful when exact transitions cannot be directly observed, but other observable data indirectly provide information about the unobservable parameters, given the model structure. The main objective of calibration is to 'reverse engineer' the model of interest; essentially, to find input parameters (e.g., TPs) so that the model predicts a known outcome (or outcomes) informed by existing data (e.g. long-term survival).