Article:

Salleh, S. orcid.org/0000-0001-9661-7125, Thokala, P., Brennan, A. et al. (2 more authors) (2017) Discrete Event Simulation-Based Resource Modelling in Health Technology Assessment. Pharmacoeconomics, 35. pp. 989-1006. ISSN 1170-7690

https://doi.org/10.1007/s40273-017-0533-1

Discrete Event Simulation based Resource Modelling in Health Technology Assessment: A Systematic Review

Syed Salleh*, Praveen Thokala, Alan Brennan, Ruby Hughes, Simon Dixon
School of Health and Related Research (ScHARR), University of Sheffield, Sheffield, United Kingdom
*Corresponding author contacts: SSAbdulRahman1@sheffield.ac.uk, +447578222898

Abstract

OBJECTIVE: To conduct a systematic review of published research on the use of discrete event simulation (DES) for resource modelling in health technology assessment (HTA). Resource modelling (RM) is broadly defined as incorporating and measuring effects of constraints on physical resources (e.g. beds, doctors, nurses) in HTA models.

METHODS: Systematic literature searches were conducted in academic databases (JSTOR, SAGE, SPRINGER, SCOPUS, IEEE, Science Direct, PUBMED, EMBASE) and grey literature (Google Scholar, NHS journal library), enhanced by manual searchers (i.e. reference list checking, citation searching and hand searching techniques).

RESULTS: The search strategy yielded 4,117 potentially relevant citations. Following the screening and manual searches, 10 articles were included. Reviewing these articles provided insights into the applications of RM: firstly, different types of economic analyses, model settings, RM and cost-effectiveness analysis (CEA) outcomes were identified. Secondly, variation in the characteristics of the constraints such as types and nature of constraints, sources of data for the constraints were identified. Thirdly, it was found that including the effects of constraints caused the CEA results to change in these articles.

CONCLUSION: The review found that DES proved to be an effective technique for RM but there were only a small number of studies applied in HTA. However, these studies showed the important consequences of modelling physical constraints and point to the need for a framework to be developed to guide future applications of this approach. (230 words)

Keywords
Discrete event simulation, resource modelling, health technology assessment, simulation modelling

Key Points for Decision Makers

- Economic evaluation studies in health care typically ignore the short-term constraints on physical resources (e.g. doctors, nurses) which can lead to incorrect results.
- Discrete Event Simulation is an effective tool for modelling the effects of constraints but there were only a small number of studies applied in health technology assessment (HTA).
- Further research is required to examine the possible developments for detailed modelling of the resource constraints in HTA.

1. Introduction

Economic evaluation in health care typically takes the long-term perspective, where all inputs are assumed to be unconstrained, i.e. the resources required by the new technology are immediately available and deployed optimally [1]. However, there may be constraints existing within the health system that may cause unintended consequences. For example, increasing clinic throughput/demand when the number of consulting rooms is constrained may require longer clinic opening hours and the need for overtime payments or in the long-run, such changes could be made through the provision of more consultation rooms. These higher costs can lead to slow implementation as providers struggle to deliver the necessary changes at the costs suggested by the economic evaluation. It is also possible that the higher costs change the incremental cost-effectiveness ratio of a technology such that it isn't cost-effective and the implemented changes produce a negative net monetary benefit.

Within health technology assessment (HTA), these issues of implementation and feasibility are typically either ignored or captured qualitatively. It has been argued that a formal quantitative assessment of diffusion, resource use and resource constraints [2], termed 'resource modelling', is required in assisting decision makers to determine whether projected uptake is feasible. Resource modelling (RM) involves estimating the numbers of different physical resources required over time within the pathway for each intervention. Thokala *et al*. [2] broadly classify resource types into two categories: (a) single-use resources which are items that can only be used once, such as pharmaceuticals, assays for diagnostic tests and some equipment such as masks, plasters and syringes; and (b) re-usable (or multiuse) resources which are those that are occupied for a given time period, but can be redeployed such as staff (e.g. doctors, nurses, consultants, laboratory technicians, administrative personnel) and equipment (e.g. hospital beds, intensive care units, ambulances, scanners). Given our focus on physical resources (e.g. doctors, nurses, beds, etc.), monetary resource constraints are not included in our definition of RM.

Modelling can be used in understanding whether capacity constraints can meet the resource demand. For single-use resources which deplete with time, it is important to understand whether there is enough capacity for the entire target population. This can be assessed with traditional HTA modelling methods (such as decision trees or Markov models) by linking the health state to resource use in order to estimate the overall resources required. For re-usable (or occupied) resources, it is important to understand the fluctuation in the resource availability to estimate whether there is enough capacity to meet the rate of demand (e.g. arrival rate of patients) and their time of occupancy (e.g. length of stay in hospitalisation). However, traditional HTA modelling methods are not suitable for modelling detailed resource usage over time [3], and there is a need for using advanced simulation techniques such as discrete event simulation (DES).

DES is a flexible modelling technique that can model the behaviour of a complex system using a sequence of well-defined events, focusing on individual entities (e.g. patients) moving through the system, and the changes in the states of the entities at discrete time points [3,4]. Importantly, DES provides the capability to model resource constraints explicitly [5,6,7]. It is therefore a useful tool for modelling re-usable (or occupied) resource constraints, as proven in its widespread use in other sectors such as operational research, engineering and scheduling [8-11]. The use of DES is also gaining momentum within the field of HTA itself [12-14]. There has been a recent systematic review on the use of DES for HTA [15], which identified 42 relevant studies. However, they excluded the studies that modelled capacity constraints.

To our knowledge, no study has attempted to review rigorously and systematically past applications of DES for constrained resource modelling in HTA. In this review, our aim was to systematically identify the economic evaluation studies using DES while accounting for (physical) resource constraints. This paper provides an overview of the methods and the results of the systematic review. The next section presents the methods used for identifying the relevant studies. Section 3 presents the synthesis of the studies at an overview level and a more detailed analysis. Section 4 presents a discussion of the key issues identified.

2. Methods

2.1 Literature searches

To identify relevant articles, a systematic literature search was conducted in 8 academic databases (JSTOR, SAGE, SPRINGER, SCOPUS, IEEE, Science Direct, PUBMED, EMBASE) and 2 other sources for grey literature (Google Scholar, NHS journal library) up till May 2017. Based on the definition of RM proposed by Thokala and colleagues [2] - the quantitative assessment of technology diffusion curves, their related resource requirements and their capacity constraints - building block techniques [16] were used to identify a list of keywords related to DES based RM in HTA, and to develop the search strategies. The final search strategies used in this review are presented in *Appendix 1*. These searches were supplemented with manual searches using the reference list checking, citation searching and hand searching techniques [17, 18].

2.2 Study selection

Two reviewers screened at the title and abstract level all articles found using the search strategy, after removing duplicates. Full texts of remaining articles were critically assessed and included if both reviewers found them relevant. The appraisal was carried out based on the following inclusion criteria: selected articles (a) reported the application of simulation based RM in HTA (i.e. measuring the effects of constraints on physical resources, when conducting budget impact analysis (BIA) or CEA) using the DES technique, and (b) were written in English language. Studies were excluded if they were not related to HTA, did not use DES, were reviews of other studies or did not assess the effects of constraints on physical resources.

2.3 Data extraction and Analysis

While reviewing the included articles, data were extracted as follows: study background (e.g. case study, applied HTA technique, data source for simulation modelling), and details of the constraints (e.g. the type of constraint, nature of the constraints). Thus, the data extracted from the articles were synthesised into two separate sections – an overview of the included articles and a detailed description of the constraints. The information gathered from these syntheses was used to identify the common themes and outcomes for discussion in this review.

3. Results

3.1 Searches and sifting

The search yielded 4,117 potentially relevant citations. After elimination by title and duplicates, 90 articles were retained for screening. The first level of screening excluded all irrelevant articles by scanning the abstracts, which led to the exclusion of 63 articles – 13 articles were not related to HTA, 34 did not use DES and 16 were reviews of other studies.

The second level of screening consisted of a full text assessment of the remaining 27 articles, which led to the exclusion of 18 articles – 2 articles were not related to HTA, 1 article was a review and 15 did not assess the effects of constraints on physical resources. The remaining 9 articles were retained and supplemented with manual searches, using reference list checking, citation searching and hand searching techniques, which identified one further article. The results from the sifting are presented visually as a PRISMA diagram in *Figure 1*.

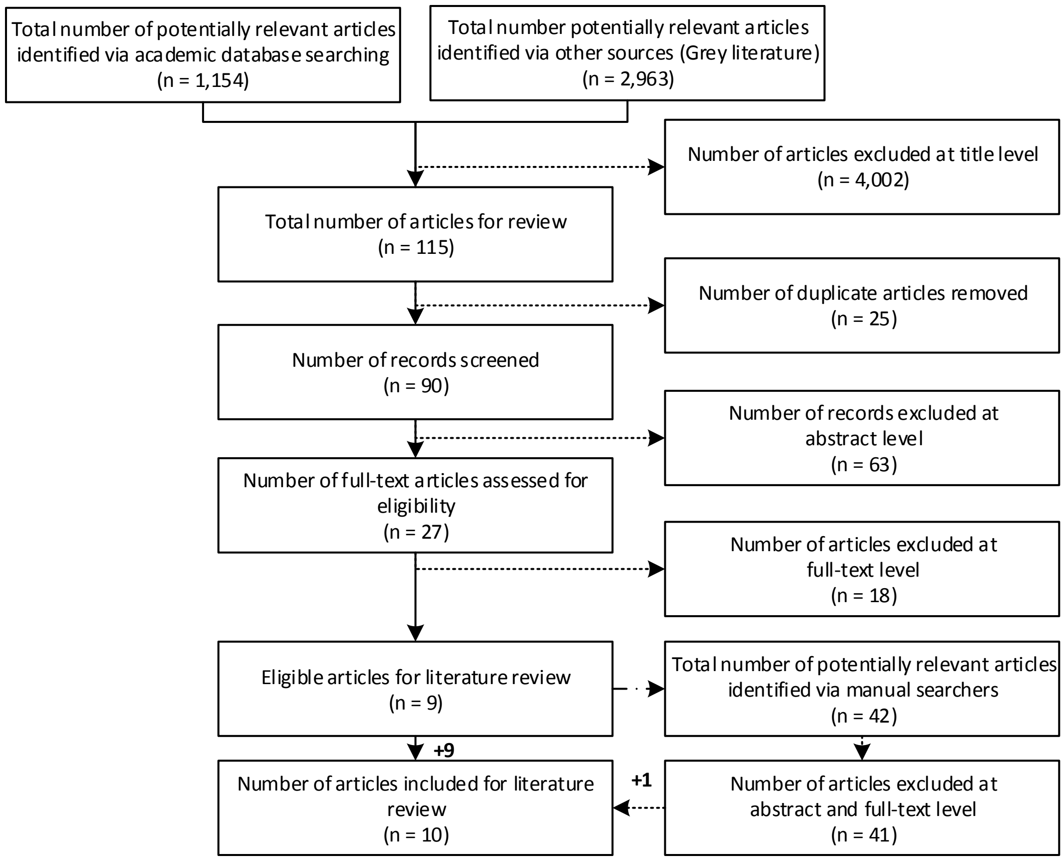

Figure 1: PRISMA diagram

3.2 Data extraction

3.2.1 Overview of the included articles

Table 1 provides a general overview of the 10 articles included in this review. Notwithstanding the diversity of applications, there are five key themes that can be identified from this overview: type of analysis performed, model setting, type of CEA outcomes, type of intervention and outcomes estimated from RM.

Table 1: The 10 articles included

No.	Study	Case study	Intervention assessed	Type of analysis	Setting	Outcomes for CEA	Type of intervention	Outcome of RM
1	[19]	Laparoscopic Cholecystectomy surgery for Anaesthesia care.	Strategy protocols for the SOR versus ORF.	Cost-effectiveness analysis	Organisational level	• Cost / Patient throughput Short-term outcomes (10h/d)	System changes: reorganizing surgical practices.	Total flow time, waiting time, patient throughput, resource utilisation
2	[20]	General surgery, gynaecology, and urology surgery for Anaesthesia	Strategy protocols for the SOR versus ORF + perioperative-staffing system (i.e. 5 cost allocation scenarios,	Cost-effectiveness analysis	Organisational level	• Cost / Patient throughput Short-term: outcomes (10h/d)	System changes: reorganizing surgical practices	Total flow time, waiting time, patient throughput, resource

		care.	consisting of 3 phases).				and workflow.	utilisation
3	[21]	Stent treatment for coronary heart disease.	Drug-eluting stent versus bare-metal stent treatments for 4 patient groups.	Cost-effectiveness analysis & budget impact analysis	Organisational level	• Cost / QALYs Long-term outcomes (7yrs)	System changes: treatment allocation scenarios.	Waiting time, patient throughput, resource utilisation, queue length
4	[22]	Glaucoma clinical service.	Alternative treatments (i.e. medical, laser or surgical techniques) and management strategies (i.e. follow-up visit times, booking cycle length).	Cost-effectiveness analysis	Organisational level	• Cost / QALYs Long-term outcomes (5yrs)	System changes: alternative follow-up times, treatment pathways and booking cycles.	Waiting time
5	[23]	Orthopaedic services (OSPC and TOMS).	Alternative scenarios for delivering increase semi- and non-urgent orthopaedic outpatient services OSPC (without additional surgical capacity) and TOMS (with and without additional surgical capacity) by an additional 25-125 patients per month.	Cost-effectiveness analysis	Organisational level	• Cost / QALYs Long-term outcomes (5.25yrs)	System changes: reorganizing orthopaedic service.	Waiting time, patient throughput, resource utilisation, queue length
6	[24]	Treatment and further prevention of CHD. **Note**: Refer to [25] for model details.	Increasing level of alternative drugs used for secondary prevention of CHD (i.e. statins, aspirin, beta blockers, angiotensin converting enzyme inhibitors).	Cost-effectiveness analysis	National level	• Cost / Lys Long-term outcomes (20yrs)	New treatment: increasing the uptake of one drug, while the comparators remain constant.	Waiting time, queue length
7	[26]	Ultrasound screening for the DDH in primary care.	Varying by locations (urban vs rural areas) and 3 screener strategies/ scenarios.	Cost-effectiveness analysis	National level	• Cost / Successful DDH detection. Short-term outcomes (3h/d)	System changes: alternative screener strategies.	Waiting time, resource utilisation
8	[27]	Robotic and laparoscopic prostatectomy treatment service for	Alternative surgical techniques (standard versus robotic) and surgical capacity per year.	Cost-effectiveness analysis	National level	• Cost / QALYs Long-term outcomes (10yrs)	New services: alternative surgical techniques.	Total flow time, waiting time, patient throughput

		localised prostate cancer.						
9	[28]	Blood collection systems (fixed and mobile sites)	Alternative configurations for capacity planning of human resources and donor appointment strategies.	Cost-effectiveness analysis	National level	• Cost / service level (i.e. percentage of entities served within a targeted waiting time). Short-term outcomes (8h/d)	System changes: reorganizing blood collection systems.	Total flow time, waiting time, waiting probability, patient throughput, resource utilisation, queue length, probability of abandonment, service level
10	[29]	Orthopaedic service **Not**e: 3 DES models are developed. 1 with DQ and 2 without).	UC versus OSPC.	Cost-effectiveness analysis	National level	• Cost / QALYs Long-term outcomes (5.25yrs)	System changes: reorganizing orthopaedic service.	Total flow time, waiting time, queue length

Abbreviations: *SOR, standard operating room; ORF, operating room of the future; QALY, quality-adjusted life year; OSPC, orthopaedic physiotherapy screening clinic and multidisciplinary treatment service; TOMS, traditional orthopaedic medical services; CHD, coronary heart disease; DDH, developmental dysplasia of the hip; DQ, dynamic queuing; UC, Usual orthopaedic care.*

RM in HTA has been performed at different settings with two main categories emerging – organisational level (i.e. RM performed at a single location) and national level (i.e. RM performed at an aggregate level). Out of the 10 articles, 5 studies applied RM at an organisational level [19-23] and 5 at a national level [24, 26-29].

Regarding the type of analyses, out of the 10 articles identified in the systematic review, only 1 article [21] includes RM aspects (i.e. resource constraints) in both CEA and BIA models, while the rest of the articles focus solely on CEA. There is a diversity of outcomes assessed alongside traditional CEA (defined as those that include QALYs, disability-adjusted life year (DALYs) or life-year saved (LYs) [30]). As presented in *table 2*, the majority of the studies solely reported CEA, either by measuring the effects using QALYs [21-23, 27, 29] or LYs [24], while the rest of the studies focused on broader outcomes such as patient throughput per day [19, 20]. Out of the 10 articles identified, only 4 of these studies assessed short-term outcomes (i.e. maximum number of cases treatable in 3, 8 or 10 hours) [19, 20, 28], while the rest of the studies only focus on long-term outcomes (e.g. costs and QALYS over 5-20 years).

There are diverse types of interventions assessed within the reviewed articles. Three categories emerged – system changes (e.g. reorganizing surgical practices), new services and new treatments. Out of the 10 articles, 8 studies assessed system changes [19-23, 26, 28, 29], 1 assessed a new service that offers robotic surgery [27] and 1 assessed increasing the uptake of secondary prevention drugs [24].

Alongside the traditional economic analyses outputs, the studies also presented a set of RM outcomes (or key performance indicators). The outcomes that were observed in these studies included total flow time [19, 20, 27-29], waiting time [19-24, 26-29], waiting probability [28], patient throughput [19-21, 23, 27, 28], resource utilisation [19-21, 23, 26, 28], queue length [21, 23, 24, 28, 29], probability of abandonment [28], and service level [28]. In one study observed [22], only one RM outcome was mentioned (waiting time); the remainder of these studies reported multiple outcomes.

Table 2 presents the details of the RM constraints for the 10 articles included in this review, which include the type of constraints, nature of constraints, sources of data constraints, uncertainty around the constraints, model performance evaluation of outputs, and the effects of constraints on CEA results.

Two main categories emerge of constraints in the studies – capacity constraints (i.e. constraints on the quantity of resources) or throughput constraints. In studies that include capacity constraints, the constraints on physical resources are explicitly modelled. For example, Stahl *et al*. [19] limited the availability of resources (e.g. nurses or surgeons) by explicitly modelling the capacity constraints. However, in some studies this is not the case, the capacity constraints are not modelled overtly but rather at a higher level of abstraction indirectly using throughput measures. For example, one study [21] limited the throughput to thirty-six patients per day as a result of the constraints in the availability of resources (e.g. doctors, specialist nurses) to meet daily demand. That is, rather than modelling the capacity constraints of resources explicitly, the potential effect of resource constraints is modelled as limited patient throughput. Out of the 10 studies identified, 2 examined the effects against capacity constraints [19, 20], 6 examined the effect against throughput constraints [21-24, 27, 29] and 2 studies examined the effects of both capacity and throughput constraints [26, 28].

Table 2: Details of the resource modelling constraints presented from included articles

No.	Study	List of resource	Characteristics of the Constraints							Conclusions from the study (i.e. effects of constraints on CEA results)
			Description	Experiment (What-if analysis)	Type of constraint *1	Nature of constraint *2	Uncertainty around constraint in the model and analysis	Model performance evaluation of outputs *3	Data source for constraint	
1	[19]	ANN, OR nurse, surgeon.	• All 3 resources were constrained	The number of ANNs were varied but the other two resources were kept fixed.	Capacity	Fixed (staffing mixed between shifts and availability of OR remains constant).	DSA (increasing the inter arrival time)	Real-world comparison, face validation.	Modeller's assumption.	• Redesigning ORF changes the waiting time. • Adding additional ANN in ORF increases the cost and throughput of PC/d. • ORF system works best when schedule of patients is greater or equal to 5 patients per day, while no longer effective if the hand-off delay >15 minutes.
2	[20]	Pre-operative nurse, certified registered nurse anaesthetist, OR	• 2 out of 9 resources were constrained (OR nurse,	The availability of FTE staffs were varied for all two	Capacity	Fixed (staffing mixed between shifts remains	DSA (increasing FTE staffs).	Real-world comparison.	Modeller's assumption.	• Redesigning ORF improves patient flow by decreasing waiting time. • Increasing the FTE staffs in ORF increases the cost and throughput of PC/d.

		nurse, OR technical staff, OR support staff, OR administrator, biomedical engineering technician, ANN, post-operative nurse.	ANN)	resources.		constant).				• CEA suggest additional costs incurred by higher staffing ratios in ORF are likely to be offset by the increase in productivity. • The ICER changes when using different cost-allocation scenarios on resources.
3	[21]	Bed.	• Resource was indirectly constrained • Limiting the number of daily patients for treatment was considered as the constraint	Limiting number of stented patients treated to 36 per day.	Throughput	Fixed (maximum number of patients accepted per day remains constant).	N/A	N/A	Modeller's assumption.	• The delay time incurred in the constrained scenario. • Limiting throughput for stented patient increases the treatment cost of PC/d (i.e. adding cost for alternative dugs until stent is given), while decreasing the QALYs (i.e. additional delay increased the risk of having angina symptoms (e.g. chest pain, etc.)). • The results of the ICERs were found dominated for most of the constraint scenarios.
4	[22]	Doctors, administration officers, specialist nurse, registrar.	• Resources were indirectly constrained • Limiting the number of daily patients for treatment considered as the constraint	Limiting number stented patients using appointment scheduling.	Throughput	Time variant (different number of patients accepted per day).	N/A	Model calibration.	Modeller's assumption, patient logs.	• Delaying patients' throughput for treatment decreases the cost (i.e. extending review time delay from 4-6 months to 1 year) and QALYs (i.e. additional delay lead to the deterioration of visual field). • The ICER changes of having to delay the patients' throughput for delivering health services (extending booking cycle and follow-up times).
5	[23]	Orthopaedic specialist,	• Resources were indirectly	Increasing or decreasing the maximum	Throughput	Fixed (maximum number of	DSA (increasing or decreasing	N/A	Modeller's assumption,	• As the maximum capacity of orthopaedic services increases, the number of patients receiving an

		physiotherapists .	constrained • Limiting the number of patients per month considered as the constraint	throughput of 25, 50, 75, 100 or 125 patients receiving orthopaedic services (OSPC or TOMS) per month.		patients accepted per month remains constant).	patients' throughput).		medical records.	initial assessment also increases, unless the supply of resources exceeds the demand. • The cost increases in line with increasing maximum throughput. • The waiting time and ICERs changes when experimenting with different patients' throughput.
6	[24]	N/A	• Resource was indirectly constrained • Limiting the number of patients per year considered as the constraint	Limiting the capacity for angiograms, bypass graft and angioplasty per year, per million population (Cooper *et al.*, 2002:263).	Throughput	Fixed (maximum number of patients accepted per year remains constant).	N/A	Real-world comparison.	Literature review.	• Limiting throughput of patients for treatment decreases the measure of LYs (i.e. due to the increase of death rate, which relates to the occurrence, type and speed of treatment for MI).
7	[26]	Paediatric physicians, nurses, radiographic technicians, ultrasound machine.	• All 4 resources were constrained • Limiting the number of children/ patient per batch in the rural and urban areas per year	• The availability of all four resources were varied. • Limiting the batch of 100 patients in urban, and 120 patients in rural areas per year.	Capacity and throughput	Fixed (staffing mixed between shifts and availability of machine, and the number of patients accepted per year remains constant).	N/A	Real-world comparison, face validation.	Modeller's assumption.	• Lower participation leads to an increase idle cost of the resources required for screening. • Adding an additional ultrasound machine decreases the waiting time, while increases the number of successful DDH detection. • The ICER indicated that applying the ultrasound screening in current infant health care would be dominant in the urban area if an additional ultrasound machine is added and the screening is organised either by physicians or nurses (scenario 3).
8	[27]	Surgeons,	• Resource	Limiting	Throughput	Fixed	DSA	Face	Modeller's	• The excess cost per case for

		operating theatre, robotic equipment, laparoscopic equipment.	was indirectly constrained •Limiting the number of surgical procedures performed per year considered as the constraint	number procedures performed by 50, 100, 150 or 200 per year.		(staffing mixed between shifts and availability of equipment remains constant).	(increasing or decreasing surgical capacity).	validation.	assumption.	robotic prostatectomy treatment can be reduced by maintaining a high throughput of 100-150 cases in each centre per year. • The ICER indicated robotic surgery being dominant in large centres that manage ≥ 200 cases per year.
9	[28]	•**Fixed site**: plasma and platelet separators •**Mobile site**: mobile blood collection unit. •**Both sites**: physicians, nurses, secretaries, WBC devices and beds.	•All 7 resources were constrained in the fixed site •5 out of 6 resources were constrained in the mobile site (physicians, nurses, secretaries, WBC devices, beds) •Limiting the number of appointment considered as the constraint in the fixed site	•The availability of physicians and nurses were varied in the fixed site but the other five resources were kept fixed. •The availability of physicians and secretaries were varied in the mobile site but the other three resources were kept fixed.	Capacity and throughput	Time variant (**Fixed site**: different appointment strategies; **Both sites**: different staffing mixed were configured using working shift).	DSA (**Fixed site** - different appointment strategies; **Both sites** - increasing or decreasing availability of resources).	Real-world comparison.	Modeller's assumption.	• The improvement in the service level of the fixed site depends not only on adequate planning of human resources, but also appointment strategies. This is for achieving a higher service level with fewer human resources and costs spend. • The probability of abandonment and waiting time decreases as the number of physicians were increased in the mobile site.
10	[29]	Orthopaedic specialist.	•Resource was indirectly constrained •Limiting the	N/A	Throughput	Time variant (patient throughput for assessment	N/A	N/A	Modeller's assumption, medical records.	• The constraint causes the outcomes for CEA to change. • The DQ model projected the highest ICER, when compared with models without it.

			number of patients accessible for the initial orthopaedic assessment and surgery in the DQ model were considered as the constraint			and surgery were dynamically generated).					• Queuing time increases for DQ model, as the demand increases.

*[*1] **Type of constraint:** (1) Capacity: Limiting the number of resource providing service/intervention (e.g. only 3 doctors are supplied in vaccination service); (2) Throughput: Limiting the number of patient access for service/intervention due to the limited supply of resources (e.g. 20 patients are accessible for vaccination service per day due to the limited supply of resources provided). [*2] **Nature of constraint:** (1) Fixed: Available physical resources or patient throughput is fixed (e.g. 50 beds or fixed number of patients accepted for treatment per day); (2) Time variant: Available physical resources or patient throughput changes with time (e.g. 10 additional beds are added to the night shift, the number of blood pack used for blood transfusion decreases as it is being used through time or different number of patients accepted for treatment per day). [*3] **Model performance evaluation technique:** (1) Real-world comparison: comparing the simulated outcomes against the reality (e.g. historical and/or observed data); (2) Face validation: validate the outcomes with an expert; (3) Model calibration: adjusting the parameter value in order to minimize the difference between the simulated and reference data. **Abbreviations:** ANN, anaesthesiologist; OR, operating room; DSA: deterministic sensitivity analyses; SU: structural uncertainty; MPE: model performance evaluation; PC/d, patients cared for per day; ICER, incremental cost-effectiveness ratio; FTE, full-time equivalent; LYs, life-year saved; MI, myocardial infarction; WBC, whole blood collection.*

The constraints identified from the reviewed articles can either be fixed over time (i.e. the availability of physical resources or patient throughput being fixed throughout the model run) or time variant (i.e. the availability of physical resources or patient throughput changes with time throughout the model run, e.g. shift patterns). 7 of the 10 studies [19-21, 23, 24, 26, 27] used fixed constraint, by having identical staffing mixes between shifts and/or constant availability of resources. Meanwhile, the remaining 3 studies [22, 28, 29] assessed time variant constraints, by using different staffing mixes between shifts [28] and/or limiting service throughput to different number of patients accepted in different days of the week (e.g. using appointment scheduling [22]).

The majority of the reviewed studies [19-21, 26-29] used simple capacity constraints. That is, these studies neglected the realistic details such as the effects of queuing or prioritising treatment by patients' severity of illness. However, these issues were incorporated in the remaining studies by introducing impatient behaviours (balking and reneging) within the queues [23] or using priority queues [22, 24]. For example, the model developed by [22] reserved throughput capacity for urgent cases. As a result, low priority patients must wait until capacity is available to serve it (i.e. if other clinics are fully booked), which consequently effects the cost-effectiveness outcome.

The sources of data constraints used for simulation based RM in HTA in the 10 articles varied. These sources ranged from modeller's assumption (e.g. expert opinion), primary data collection (i.e. patient logs, medical records), and secondary data (e.g. literature review). The most common was modeller's assumption (n=9), followed by medical records (n=2) and patient logs (n=1). Only three models [22, 23, 29] combined multiple sources of data constraints, while the remaining focused on a single source. It should also be noted that out of the 5 national level models, only 1 [24] used national constraint data (e.g. rates published by the British Cardiovascular Society). Meanwhile, the rest of the models [26-29] gathered constraint data at an organisation level (i.e. modeller's assumption and/or reviewing medical records)

Uncertainty associated with the constraints was included in the modelling and analysis in 5 out of 10 articles. These examined uncertainty in the parameters associated with the patients inter arrival time [19], full-time equivalent (FTE) staff members [20] and service throughputs by performing deterministic sensitivity analyses [23, 27], and 1 study examined in both the staffs and service throughputs [28]. The remaining studies examined stochastic uncertainty in the results using probability sensitivity analysis (PSA) by conducting multiple simulation runs varying all parameters including those relating to constraints [21, 22, 24, 26, 29].

7 of the 10 studies [19, 20, 22, 24, 26-28] mentioned the use of model performance evaluation to determine the validity of constraint outputs. Three techniques were identified. The most common technique was real-world comparison (n=5), followed by face validation (n=3) and model calibration (n=1). 2 of these 7 studies [19, 26] used mixed techniques; the remaining used individual technique. Furthermore, out of the 3 studies [21, 23, 29] that did not perform model performance evaluation, only one [23] mentioned the reason for not evaluating the constraint output (i.e. limited availability of data for real-world comparison).

All the studies concluded that incorporating the effects of constraints changes the RM and CEA outcomes. However, the effects of the constraints are varied. Within the results of the 10 articles included in this review: 9 studies [19-23, 26-29] found changes in costs; 9 studies [19-24, 26, 28, 29] found changes in effectiveness outcomes (e.g. QALYs,); 7 studies [20-23, 26, 27, 29] indicated changes in ICER; 3 studies [19, 20, 23] indicated changes in patients' throughput; and 8 studies [19-23, 26, 28, 29] indicated changes in waiting time. Therefore, neglecting these constraints may lead to incorrect results. It should be noted that only one study [24] reported a single effect of constraint (i.e. changes in costs), with the remaining studies reported multiple effects.

4. Discussion

This paper set out to provide a comprehensive and systematic review of the studies that report on the use of DES for RM in HTA. RM in HTA is a relatively new topic [2]. So, this paper aims to provide an overview of its application as well as to suggest directions for further research.

RM is useful in situations where the CEA results are affected by constraints (as assumed by the stakeholders e.g. doctors) and/or if the new technology is reliant on change of physical resources. There are several ways in which these situations can manifest themselves and these have been highlighted using examples from reviewed articles as follows:

- Need for additional new resources when using a new technology [19, 20, 26].
- Need for specialised resources when administering the new technology [19, 20, 22, 23, 26-29].
- Having a very limited supply of resources when administering the new technology [19-24, 26-29].
- Having effect on the existing queues/waiting list when administering the new technology [21, 24].

The potential impact of constraints can be categorised into two types – process impacts and health impacts. Incorporating constraints will always result in process impacts i.e. delays, waiting lists, etc. However, these process impacts may also manifest as health impacts – e.g. if delay to certain treatments affects the rate of recovery and mortality, as in the case of thrombolysis [31]. It should be noted that an individual constraint can have more than one type of impact. Out of the 10 articles identified, the majority (n=7) focus on the process impacts such as the effect of delays on the service flow and associated costs while only three studies incorporated the possible effect on a patient's health owing to treatment delay [21, 22, 24]. These studies focus on coronary heart disease (increasing rate of angina symptoms with treatment delay; mortality correlated to the speed of treatment) and glaucoma (increasing rate of deterioration of the visual field with treatment delay). If service delay were a serious concern for a given condition [32, 33], neglecting the effect of constraints in the model may lead to changes in health outcomes (e.g. QALYs, LYs), and hence wrong cost-effectiveness results [21].

The usefulness of any model depends on the accuracy and validity of the outputs. It is obvious that the quality of data used to model the constraints has a direct impact on the quality of the model results and neglecting the quality of constraint data may lead to misleading results and conclusions in HTA. However, as observed in the studies included in the review most of the data on the constraints was based on opinion and assumption. This is typical of these types of studies given the constraints are context/setting specific making it difficult to use the data from published literature. However, the model outputs can be validated (with real world data, where available) in order to ensure the validity of the assumptions/opinions. Model performance evaluation was conducted by validating the results of the model [19, 20, 24, 26, 28] or calibrating [22] the effects of uncertainty explicitly to the data observed in real-world. If there is no real world data available, face validation can be performed with experts to ensure that the results are sensible, as reported in some of the studies [19, 26, 27]. However, face validity is subjective and should be treated with caution. Data can also be collected to inform the model and/or to validate the results of the model [19, 20, 24, 28].

Precise data are rarely available and there is need to account for the effects of uncertainty on the constraints. Sensitivity analysis can help to quantify this uncertainty. Five studies [19, 20, 23, 27, 28] incorporated uncertainty in the parameters (input values, e.g. availability of resources, throughputs) relating to the constraints, using deterministic sensitivity analysis. For example, a systematic analysis was conducted in one of the studies to explore the effects on the costs, QALYs and ICERs for different patient throughputs of having surgical constraints [27]. Meanwhile, the remaining studies [21, 22, 24, 26, 29] quantify uncertainty in the results using PSA by running multiple simulation runs, varying all relevant parameters including those relating to constraints, in order to produce better estimate of mean. For example, one study used 1000 simulation runs to generate stable ICER estimates [29]. Given the exact values (parameters) are unknown and that the inclusion of constraints impacts the results, an exploration of the uncertainty around them should also be undertaken. There are existing reviews and guidelines for modelling uncertainty in HTA that can be referred to for more information [34-38].

An aspect observed in all the studies using DES based RM, as opposed to a standard HTA model relates to the model time frame [39]. Long-term modelling is needed for estimating the cost-effectiveness, in order to capture all relevant outcomes [40] while the need to understand short-term fluctuations in the resource capacity needs the model to produce outputs in the short-term (e.g. a few months). It should be noted that the short-term resource capacity can only be captured in models that explicitly model the constraints.

The models that are based on 'limited throughput' can only provide the long-term results as they cannot capture the short-term resource issues. When performing resource modelling, thus, if there is a need to model both the short-term process related outcomes as well as the long-term health (and cost) outcomes, the models need to explicitly incorporate the resource constraints in order to accurately estimate the cost-effectiveness as well as the resource issues. 4 of the 10 studies in our review used such approach and combined both of these outcomes in models with capacity constraints [19, 20, 26, 28]. For example, the model developed by Ramwadhdoebe [26] calculated the short-term cost for waiting per day of having limited resources (e.g. paediatric physicians). This process is replicated until the end of the 5 year simulation period to estimate the long-term outcomes, hence providing a better representation of the health service and cost-effectiveness outcomes. However, the remaining studies

that examined throughput constraints [21, 22, 23, 24, 27, 29] focused mainly on long-term process outcomes. These studies are not able to capture the short-term process delays as a result of micro-level interactions (of individual patients competing for scarce resources).
The studies that considered throughput constraints only implicitly addressed the constraints [21-24, 27, 29]. These studies do not model the resource usage (e.g. working patterns, utilisation) or the effects of delays in individual processes (e.g. queues, waiting lists), but rather combine everything into a composite measure (limited patient throughput) and use it as a proxy for resource constraints. There are drawbacks to this approach. Firstly, using a proxy 'throughput' does not give the whole picture, and may produce different results compared to an explicit constrained model. Secondly, these analyses provide less flexibility for allowing a wider range of experiments (on physical resources) to be carried out. Thirdly, these analyses do not capture the full effects of queuing due to resource constraints. Effects of queuing such as balking (refusing to join the queue) or reneging (leaving the queue after entering) cannot be explicitly considered in the 'limited throughput' models. If any of these issues are important, then a fully constrained DES model is required to accurately determine cost-effectiveness in models; using a 'throughput' model will not capture the full effect of constraints. Whilst the throughput can be varied to proxy short-term constraints (e.g. different number of patients accepted per day [22]), it does not provide the same level of detail as modelling the constraints explicitly. It should be noted that DES modelling is not a prerequisite for capturing the effect of constraints via throughput measures - 'simpler' approaches that predominate in HTA such as cohort-based state transition models (such as Markov models) and decision tree models can be used for estimating overall resource requirements. Two such studies reported the use of Markov modelling to estimate the resources required in order to estimate the feasibility of the respective programmes [41, 42].

In reviewing the literature, it is clear that explicit modelling of the constraints (using capacity constraint) leads to a more realistic forecast and offers a superior forecasting capability for DES based RM, when compared to modelling throughput constraints. However, the choice of approach depends on the model aim. For example, using throughput constraint may well be adequate in some situations to produce a generalisable conclusion. Also, as observed in one study [21], whilst applying RM for HTA provides a better representation of reality (real-world capacity restrictions), this may increase the modelling and running time of the model. The value of complex DES modelling should be assessed within the context of the decision-making process and the extent to which the constraints could affect the cost-effectiveness as identified within the project scoping (e.g. discussion with the stakeholders). This review has focused only on DES. However, it is believed that there could be other techniques that can be used for RM in HTA such as system dynamics (SD) or agent based modelling (ABM) [43] or cohort modelling (Markov models or decision trees) [44].

The use of resource modelling also has links to the concept of value of implementation analysis. The expected value of implementation (EVImp) is the net monetary benefit associated with increasing the uptake of a cost-effective technology [45]. EVImp provides a starting point for further research that looks at whether it is cost-effective to invest in ways of speeding up the uptake of cost-effective technology. Only a small number of studies have estimated EVImp [46-47] and these have been based on changes in overall uptake (i.e. implicitly modelling the constraints have limited uptake). Resource modelling, could provide another way of looking at EVImp by directly modelling the additional costs required to loosen a constraint on throughput and assessing whether the revised ICER still falls below the funding threshold.

Other than the few studies that were identified in our review [19, 20, 26, 28], most HTA studies do not address the effects of constraints explicitly when developing a patient-level model [15]. The reviewed studies showed that DES is an effective tool to assess the effects of constraints, but given the small number of studies found in our review, there is a gap in understanding on how DES can be used for considering explicit constraints when performing RM in HTA. On the other hand, resource modelling is a common feature in traditional operational research (OR) studies. A recent umbrella review [11] identified twelve systematic reviews on the application of DES models in health care, with majority of the studies included in these reviews relating to capacity planning. Whilst these OR studies model the process efficiency (i.e. effect of resource constraints on process outcomes such as waiting times, queues, etc.) in detail, the health outcomes are not captured at all. This is in contrast with that of HTA, where the long-term outcomes are modelled in detail but the resource constraints are ignored. There is a need to consolidate the practices from both OR and HTA fields in order to establish good practice when estimating long-term outcomes taking resource constraints into account.

5. Conclusion

In this paper, a systematic review of articles on simulation based RM in HTA, using discrete event simulation models is reported. The studies showed that DES proved to be an effective technique for RM in HTA and the constraints can be important and can affect the cost-effectiveness results. However, the field is still in its infancy with issues still surrounding which modelling technique to use, how exactly to incorporate resource constraints and their effects on the system, what data sources should be used to use model resource constraints, how to incorporate uncertainty in these resource models, and how to validate the outputs from these models. Further research should be undertaken to examine the possible developments for detailed modelling of the resource constraints in HTA in order that more robust and valuable outputs from such analyses are produced.

Data availability statement

All data generated or analysed during this study are included in this published.

Acknowledgements

We thank all of those who contributed to this work. In addition, the corresponding author (SS) would like to thank Majlis Amanah Rakyat (MARA) for sponsoring this project. The study idea originated from PT and SD, which was further developed in discussions with SS, and finalised in communication with all authors. SS coordinated the data collection and interpretation, which was agreed among all authors. SS and PT wrote the initial draft, with all authors contributing to the submitted version and also revising the manuscript based on reviewers' comments. SS is the overall guarantor for the manuscript.

Compliance with ethical standards

Funding This work was supported by the Majlis Amanah Rakyat.

Conflict of interest Syed Salleh declares that he has no conflicts of interest. Praveen Thokala declares that he has no conflicts of interest. Alan Brennan declares that he has no conflicts of interest. Ruby Hughes declares that she has no conflicts of interest. Simon Dixon declares that he has no conflicts of interest.

References

1. Weinstein M. Recommendations of the Panel on Cost-effectiveness in Health and Medicine. The Journal of the American Medical Association. 1996;276(15):1253-1258.
2. Thokala P, Dixon S, Jahn B. Resource Modelling: The Missing Piece of the HTA Jigsaw?. PharmacoEconomics. 2015;33(3): 193-203.
3. Marshall D, Burgos-Liz L, IJzerman M, Osgood N, Padula W, Higashi M, Wong P, Pasupathy K, Crown W. Applying Dynamic Simulation Modeling Methods in Health Care Delivery Research—The SIMULATE Checklist: Report of the ISPOR Simulation Modeling Emerging Good Practices Task Force. Value in Health. 2015;18(1):5-16.
4. Marshall D, Burgos-Liz L, IJzerman M, Crown W, Padula W, Wong P, Pasupathy K, Higashi M, Osgood N. Selecting a Dynamic Simulation Modeling Method for Health Care Delivery Research—Part 2: Report of the ISPOR Dynamic Simulation Modeling Emerging Good Practices Task Force. Value in Health. 2015;18(2):147-160.
5. Karnon J, Stahl J, Brennan A, Caro J, Mar J, Moller J. Modeling Using Discrete Event Simulation: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-4. Medical Decision Making. 2012;32(5):701-711.
6. Caro J, Briggs A, Siebert U, Kuntz K. Modeling Good Research Practices—Overview: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force-1. Medical Decision Making. 2012;32(5):667-677.
7. Standfield L, Comans T, Scuffham P. Markov modeling and discrete event simulation in health care: a systematic comparison. International Journal of Technology Assessment in Health Care. 2014;30(02):165-172.
8. Gunal M, Pidd M. Discrete event simulation for performance modelling in health care: a review of the literature. Journal of Simulation. 2010;4(1):42-51.
9. Konig M, Beibert U, Steinhauer D, Bargstadt H. Constraint-based simulation of outfitting processes in shipbuilding and civil engineering. In: Eurosim Congress on Modelling and Simulation. 2007:1-11. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.462.6589. Accessed 10 Aug 2016.
10. Cardoen B, Demeulemeester E, Belien J. Operating room planning and scheduling: A literature review. European Journal of Operational Research. 2010;201(3):921-932.
11. Salleh S, Thokala P, Brennan A, Hughes R, Booth A. Simulation Modelling in Healthcare: An Umbrella Review of Systematic Literature Reviews. PharmacoEconomics. 2017:1-13.
12. Caro J, Moller J, Getsios D. Discrete Event Simulation: The Preferred Technique for Health Economic Evaluations?. Value in Health. 2010;13(8): 1056-1060.
13. Caro J, Moller J. Advantages and disadvantages of discrete-event simulation for health economic analyses. Expert Review of Pharmacoeconomics & Outcomes Research. 2016;16(3):327-329.
14. Caro J, Moller J, Karnon J, Stahl J, Ishak J. Discrete event simulation for health technology assessment. Boca Raton: CRC Press; 2016. pp.227. ISBN 978-1-4822-1824-4
15. Karnon J, Haji Ali Afzali H. When to Use Discrete Event Simulation (DES) for the Economic Evaluation of Health Technologies? A Review and Critique of the Costs and Benefits of DES. PharmacoEconomics. 2014;32(6):547-558.

16. Booth A. Unpacking your literature search toolbox: on search styles and tactics. Health Information & Libraries Journal. 2008;25(4):313-317.
17. Gough D, Oliver S, Thomas J. An introduction to systematic reviews. London: SAGE; 2012. pp.125-126. ISBN 9781-4-4625-8705
18. Booth A, Papaioannou D, Sutton A. Systematic approaches to a successful literature review. London: SAGE; 2012. pp.78-79. ISBN 9780-8-5702-1359
19. Stahl J. Reorganizing the System of Care Surrounding Laparoscopic Surgery: A Cost-Effectiveness Analysis Using Discrete-Event Simulation. Medical Decision Making. 2004;24(5):461-471.
20. Stahl J, Sandberg W, Daily B, Wiklund R, Egan M, Goldman J, Isaacson K, Gazelle S, Rattner D. Reorganizing patient care and workflow in the operating room: a cost-effectiveness study. Surgery. 2006;139(6):717-728.
21. Jahn B, Pfeiffer K, Theurl E, Tarride J, Goeree R. Capacity Constraints and Cost-Effectiveness: A Discrete Event Simulation for Drug-Eluting Stents. Medical Decision Making. 2009;30(1):16-28.
22. Crane G, Kymes S, Hiller J, Casson R, Martin A, Karnon J. Accounting for Costs, QALYs, and Capacity Constraints: Using Discrete-Event Simulation to Evaluate Alternative Service Delivery and Organizational Scenarios for Hospital-Based Glaucoma Services. Medical Decision Making. 2013;33(8):986-997.
23. Standfield L, Comans T, Raymer M, O'Leary S, Moretto N, Scuffham P. The Efficiency of Increasing the Capacity of Physiotherapy Screening Clinics or Traditional Medical Services to Address Unmet Demand in Orthopaedic Outpatients: A Practical Application of Discrete Event Simulation with Dynamic Queuing. Appl Health Econ Health Policy. 2016;14(4):479-91.
24. Cooper K, Davies R, Raftery J, Roderick P. Use of a coronary heart disease simulation model to evaluate the costs and effectiveness of drugs for the prevention of heart disease. J Oper Res Soc. 2007;59(9):1173-1181.
25. Cooper K, Davies R, Roderick P, Chase D, Raftery J. The Development of a Simulation Model of the Treatment of Coronary Heart Disease. Health Care Management Science. 2002;5(4):259-267.
26. Ramwadhdoebe S. Screening for developmental dysplasia of the hip in primary care. Implementation by simulation. Ph.D. Utrecht University. 2010. pp.77-100. ISBN: 978-90-393-5449-0.
27. Ramsay C, Pickard R, Robertson C, Close A, Vale L, Armstrong N, Barocas D, Eden C, Fraser C, Gurung T, Jenkinson D, Jia X, Lam T, Mowatt G, Neal D, Robinson M, Royle J, Rushton S, Sharma P, Shirley M, Soomro N. Systematic review and economic modelling of the relative clinical benefit and cost-effectiveness of laparoscopic surgery and robotic surgery for removal of the prostate in men with localised prostate cancer. Health Technology Assessment. 2012;16(41).
28. Alfonso E, Xie X, Augusto V, Garraud O. Modelling and simulation of blood collection systems: improvement of human resources allocation for better cost-effectiveness and reduction of candidate donor abandonment. Vox Sanguinis. 2012;104(3):225-233.
29. Standfield L, Comans T, Scuffham P. An empirical comparison of Markov cohort modeling and discrete event simulation in a capacity-constrained health care setting. The European Journal of Health Economics. 2015;18(1):33-47.
30. Robberstad B. QALYs vs DALYs vs LYs gained: What are the differences, and what difference do they make for health care priority setting?. Global health research. 2005;15(2):183-191.

31. De Luca G, Suryapranata H, Ottervanger J, Antman E. Time Delay to Treatment and Mortality in Primary Angioplasty for Acute Myocardial Infarction: Every Minute of Delay Counts. Circulation. 2004;109(10): 1223-1225.
32. Siciliani L, Borowitz M, Moran V. Waiting time policies in the health sector. Paris: OECD; 2013. p.238. ISBN 978-92-64-17908-0.
33. Kate S, Richard L, Helen B, Richard S, Paul W. Reducing waiting times in the NHS: is lack of capacity the problem?. Clinician in Management. 2004;12(3):105-109.
34. Briggs A. Handling Uncertainty in Cost-Effectiveness Models. PharmacoEconomics. 2000;17(5): 479-500.
35. Briggs A, Gray A. Handling uncertainty when performing economic evaluation of healthcare interventions. Health Technology Assessment. 1999;3(2):1-134.
36. Andronis L, Barton P, Bryan S. Sensitivity analysis in economic evaluation: an audit of NICE current practice and a review of its use and value in decision-making. Health Technology Assessment. 2009;13(29):1-61.
37. Briggs A, Weinstein M, Fenwick E, Karnon J, Sculpher M, Paltiel A. Model Parameter Estimation and Uncertainty Analysis: A Report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group-6. Medical Decision Making. 2012;32(5):722-732.
38. Bojke L, Claxton K, Sculpher M, Palmer S. Characterizing Structural Uncertainty in Decision Analytic Models: A Review and Application of Methods. Value in Health. 2009;12(5):739-749.
39. Davis S, Stevenson M, Tappenden P, Wailoo A. Cost-effectiveness modeling using patient-level simulation. NICE DSU technical support document 15. Sheffield: The Decision Support Unit (DSU). 2014: 7. https://www.ncbi.nlm.nih.gov/pubmed/27466644. Accessed 11 May 2015.
40. Mathes T, Jacobs E, Morfeld J, Pieper D. Methods of international health technology assessment agencies for economic evaluations- a comparative analysis. BMC Health Services Research. 2013;13(1):371.
41. Sharp L, Tilson L, Whyte S, Ceilleachair A, Walsh C, Usher C et al. Using resource modelling to inform decision making and service planning: the case of colorectal cancer screening in Ireland. BMC Health Services Research. 2013;13(1).
42. Miquel-Cases A, Steuten L, Rigter L, van Harten W. Cost-effectiveness and resource use of implementing MRI-guided NACT in ER-positive/HER2-negative breast cancers in The Netherlands. BMC Cancer. 2016;16(1).
43. Chhatwal J, He T. Economic Evaluations with Agent-Based Modelling: An Introduction. PharmacoEconomics. 2015;33(5):423-433.
44. Briggs A, Sculpher M. An Introduction to Markov Modelling for Economic Evaluation. PharmacoEconomics. 1998;13(4):397-409.
45. Fenwick E, Claxton K, Sculpher M. The Value of Implementation and the Value of Information: Combined and Uneven Development. Medical Decision Making. 2007;28(1):21-32.
46. Whyte S, Dixon S, Faria R, Walker S, Palmer S, Sculpher M, Radford S. Estimating the Cost-Effectiveness of Implementation: Is Sufficient Evidence Available? Value in Health. 2016;19(2):138-144.

47. Faria R, Walker S, Whyte S, Dixon S, Palmer S, Sculpher M. How to Invest in Getting Cost-effective Technologies into Practice? A Framework for Value of Implementation Analysis Applied to Novel Oral Anticoagulants. Medical Decision Making. 2016;37(2):148-161.