Received: 25 December 2019 | Revised: 18 December 2020 | Accepted: 19 December 2020

DOI: 10.1002/sim.8882

RESEARCH ARTICLE

Statistics in Medicine WILEY

Power and sample size for multistate model analysis of longitudinal discrete outcomes in disease prevention trials

Isabelle L. Smith[1] | **Jane E. Nixon**[1] | **Linda Sharples**[2]

[1]Clinical Trials Research Unit, University of Leeds, Leeds, UK

[2]Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK

Correspondence
Isabelle L. Smith, Clinical Trials Research Unit, University of Leeds, Leeds, UK.
Email: i.l.smith@leeds.ac.uk

Funding information
Health Technology Assessment Programme, Grant/Award Number: 11/36/33; Research Trainees Coordinating Centre, Grant/Award Number: DRF-2016-09-085

For clinical trials where participants pass through a number of discrete health states resulting in longitudinal measures over time, there are several potential primary estimands for the treatment effect. Incidence or time to a particular health state are commonly used outcomes but the choice of health state may not be obvious and these estimands do not make full use of the longitudinal assessments. Multistate models have been developed for some diseases and conditions with the purpose of understanding their natural history and have been used for secondary analysis to understand mechanisms of action of treatments. There is little published on the use of multistate models as the primary analysis method and potential implications on design features, such as assessment schedules. We illustrate methods via analysis of data from a motivating example; a Phase III clinical trial of pressure ulcer prevention strategies. We clarify some of the possible estimands that might be considered and we show, via a simulation study, that under some circumstances the sample size could be reduced by half using a multistate model based analysis, without adversely affecting the power of the trial.

KEYWORDS
clinical trial design, discrete outcomes, longitudinal data, multistate model

1 | BACKGROUND

In randomized controlled trials (RCT) of disease prevention treatments, patients may be observed to pass through a series of health states, for example, in cancer trials progression or recurrence of disease, diagnosed by imaging scans or biomarkers, may occur before death; similarly, clinical events such as MI or stroke may precede cardiovascular death in heart disease trials.[1] In trials of chronic conditions such as hand eczema, longitudinal measures of a discrete outcome at fixed time points may demonstrate increasingly severe disease classification over time.[2] Although the final endpoint may be mortality or disease severity above a certain threshold, the observed intermediate events and/or disease states can provide useful information on both the mechanisms of action and the overall efficacy of prevention treatments.

Multistate models (MSM) are structures that represent different disease categories (states) and movement of patients between these disease categories (transitions). They are convenient representations of diseases that can be classified into distinct categories, with clear definitions, and where onset, progression, and regression of the disease correspond to transitions between states in the model. MSM have been used to explore the natural history of diseases and conditions as diverse as lung transplantation,[3] cardiovascular diseases,[4] chronic myeloid leukemia,[5] colon cancer,[6] and psoriatic arthritis.[7] Moreover, they have been applied to large disease cohorts and registry data, primarily for prediction of patient prognosis.[8,9] Both general statistical methods and software to fit models to observed data are available.[10-12]

MSM have been used in secondary analysis of RCT data to better understand the mechanisms underlying primary analysis results, for example, the illness-death model to explore disease recurrence/progression patterns in chronic myeloid leukemia.[5] They are also commonly used in health economics decision models.[13] However, despite recent research concluding that they are a promising tool for use in clinical trials,[14] to our knowledge, there is only one publication assessing their potential in terms of type I error and power via a simulation study.[15] The authors used data from trials in the stroke setting and considered multistate models with 4, 5, 6, and 7 states compared with repeated logistic regression. The authors conclude that when the treatment effect is the same for all transitions bar one, multistate models provide increased power compared with repeated logistic regression; however, when the treatment effects differ across all transitions, repeated logistic regression models are more powerful. For multistate models, the overall effect of treatment on the disease process was tested using a likelihood ratio test rather than considering the treatment effect on specific disease transitions.

At the design stage of a disease prevention trial, investigators need to make a number of interrelated decisions about the patient population, overall size of the trial, the length of patient follow-up, and the intervals between patient assessments. This article aims to show how MSM can contribute to these decisions.

First, MSM of disease processes, fitted to disease cohorts and registry data can provide insight into the number of patients who might be suitable for participation in a clinical trial and the number of events of a specific type that will be observed in a stated timescale.[9,16] As an example, Le Rademacher et al compared a 3-state MSM and time-dependent Cox models in cancer clinical trials using simulation, highlighting the insight that MSM provide into the disease process and corresponding treatment effects.[17]

Second, MSM can use longitudinal data efficiently. In RCTs where the outcome requires detection of a new disease or a particular stage of disease, patient assessment often takes place at a number of fixed time points, resulting in serial measurements. Panel data of this type, for which only snapshots of the underlying disease process are obtained, are interval censored, in that changes in health status can occur at some time point between assessments, and only part of the (latent) disease process is observed.[12] In cancer trials that use progression assessed via imaging or other tests, ignoring this interval censoring has been shown to result in sample size estimates that are up to 7.2% lower than required for the stated power.[18] Therefore, sample size estimation should consider interval censoring at the design stage, and MSM together with parametric transition models provide an appropriate method for accommodating it.[10-12]

Third, in practice, due to administrative and patient-related events, assessments are not necessarily conducted at the same time point for all participants, and time intervals between assessments may vary.[19] Continuous-time MSM can estimate unbiased transition rates and treatment effects when assessment time intervals vary, provided that the measurements themselves are independent of the fact that a measurement was taken.[19]

Fourth, using intermediate health states in MSM may result is smaller trials. Incidence of death or severe disease may be easier to define and is often the estimand of choice in RCTs, but such endpoints may occur rarely, resulting in the need for very large trials. The impact of using MSM to inform treatment effects and increase power of a RCT is unclear.

Fifth, MSM can be used to assess the impact of frequency of assessments. A simulation study by Zeng et al explored efficiency gains due to increasing the frequency of patient assessments in an illness death model, concluding that, in their context, the gain in power was small in comparison to increasing the sample size.[18] Of note, their model constrained the well-to-disease progression and well-to-death transition intensities to be equal, with power estimates based on the effect of treatment on time to any transition out of the well state.

Despite potential improved trial efficiency and greater understanding of treatment mechanisms for MSM, possible barriers to their use for primary analysis of RCTs have been raised. For instance, MSM have a more complicated structure than simple regression models, so that a number of estimands may be of interest, but which should be primary? Although MSM can be used to calculate traditional endpoints, such as incidence of a particular event or disease category, choice of the specific structure of the model is not necessarily clear-cut. Further, Manzini et al highlighted the need for sufficient numbers of observed transitions throughout the MSM structure and difficulties in dealing with missing data in this context.[14] Le Rademacher et al acknowledged that there are barriers to their use in practice including availability of easily accessible validated software, and interpretation of the results of multistate model analyses.[17]

This article investigates the use of MSM to analyze discrete, interval censored longitudinal data in the context of RCTs. Motivation for this work arose from trials of treatments for the prevention of pressure ulcers (PUs), which are classified on a 4-point ordinal scale from 1 to 4, with 4 the most severe category (Table 1).[21] This semiquantitative scale for PU classification and the longitudinal data arising from trials suggests that MSM would fit this situation well. As the number

TABLE 1 Adapted EPUAP/NPUAP PU Classification system[21]

State	PU assessment	Description
1	Healthy	No skin changes
2	Altered	Alterations to intact skin
3	Category/stage I	Intact skin with nonblanchable redness of a localized area usually over a bony prominence. Discoloration of the skin, warmth, edema, hardness, or pain may also be present. Darkly pigmented skin may not have visible blanching
4[a]	Category/stage II	Partial thickness loss of dermis presenting as a shallow open ulcer with a red pink wound bed, without slough. May also present as an intact or open/ruptured serum filled blister
	Category/stage III	Full thickness tissue loss. Subcutaneous fat may be visible but bone, tendon, or muscle is not exposed. Slough may be present but does not obscure the depth of tissue loss. May include undermining and tunneling
	Category/stage IV	Full thickness tissue loss with exposed bone, tendon, or muscle. Slough or eschar may be present on some parts of the wound bed. Often include undermining and tunneling

[a] Category/stage II, III, and IV are combined into a single absorbing state based on clinically meaningful disease status.

of older people in the UK and around the world increases, the population at high risk of PUs also increases, so that measures to prevent them and efficient evaluation in clinical trials are important. This is heightened by the vulnerable nature of the at-risk population. The case of PU prevention trials is used throughout this article for illustration, but the methods are general and could be applied to any trials in which discrete outcomes are measured repeatedly through time.

The remainder of the article is structured as follows: an introduction to MSM and some notation, followed by an illustration of methods applied to an existing data set from a PU prevention trial, a simulation study to examine the impact on sample size and power under different scenarios, and finishes with a discussion of the findings and implications for further research.

2 | MULTISTATE MODELS

For the purposes of this article, we will consider a 4-state MSM with three transient states (healthy, preclinical, and mild disease) and one absorbing state (severe disease) (Figure 1). For the example of PUs these states are given by rows 1-4 of Table 1. In this 4-state model, regression from more to less severe health states (ie, healing) is not allowed, so that each state represents the most severe category observed up to the current observation, but this is not a requirement of this approach.

As with survival models, assuming measurements are taken in continuous time, movement between health states is according to the hazard or intensity for each transition, which can be summarized in a transition intensity matrix $Q(t)$, where the rsth element, $q_{rs}(t)$ denotes the instantaneous probability of transition from state r to state s at time t. Suppressing dependence on time, the corresponding transition intensity matrix, Q, for the multistate model illustrated in Figure 1 is given by (1).

$$\boldsymbol{Q} = \begin{pmatrix} -q_{12} & q_{12} & 0 & 0 \\ 0 & -q_{23} & q_{23} & 0 \\ 0 & 0 & -q_{34} & q_{34} \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{1}$$

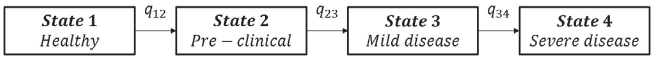


FIGURE 1 Multi-state model for 4-state disease progression

The intensity for transition $r \to s$ at time t is defined by (2)

$$q_{rs}(t) = \lim_{\delta t \to 0} \frac{Prob\,(t \leq T < t + \delta t | T \geq t, \mathcal{F}_t)}{\delta t}, \tag{2}$$

where $\mathcal{F}_t$ denotes the disease history of the patient up to time t. For a Markov model we assume that the hazard rate for the transition $r \to s$ is independent of $\mathcal{F}_t$, which greatly simplifies the model.

Time t denotes the time since the process began, which for RCT is usually the time that the patient was randomized, known as a clock forward approach.[22] Note that, in situations where time of entry and exit to each health state is known and observed, a survival function can be estimated for each transition from state entry, known as a clock reset approach.[22] The full likelihood is the product of probabilities of observed transitions between states over all individuals j and observation times l given by (3).

$$L(Q) = \prod_l \prod_j prob\left(X\left(t_{j,l+1}\right) = s | X\left(t_{jl}\right) = r\right), \tag{3}$$

where $X(t_{jl})$ denotes the state occupied by patient j at time l. Note that these probabilities are functions of the transition intensities. Estimates of transition intensities and hazard ratios for treatment effects can be estimated by maximizing the likelihood in Equation (3). However, all but the simplest multistate models are complicated and cannot be expressed in closed form. Jackson developed the *msm* package within *R* which uses eigensystem decomposition to maximize the likelihood and obtain estimates of q_{rs} assuming parametric transition time distributions.[12] Covariates may be incorporated into transition-specific regression models in line with (4)

$$q_{rs}(t) = q_{rs}(t|\boldsymbol{z}(t)) = q_{rs.0}(t) \exp\left(\boldsymbol{\beta}_{rs}^T \boldsymbol{z}(t)\right), \tag{4}$$

where $q_{rs.\,0}(t)$ denotes the baseline hazard, $\boldsymbol{\beta}_{rs}$ is a parameter vector of length k corresponding to the covariate vector $\boldsymbol{z}(t)$ also of length k. The key covariates in RCTs are the treatment contrasts and stratification variables. A common assumption is that transition intensities are constant through time although piecewise constant hazards can be used if this assumption is not valid.[23]

3 | MOTIVATING EXAMPLE

PRESSURE2 was a Phase III RCT of 2029 high-risk patients designed to compare two types of mattress.[24] To minimize misclassification, the primary outcome was the time to development of a new Category II or more severe PU.[25] The treatment phase was up to 60 days and patients were assessed twice a week for 30 days, and once a week for a further 30 days or until no longer at high risk, were discharged from hospital, withdrew, or died. A subset of 1846 patients who entered the trial without an existing Category II or more severe PU and had at least one follow-up assessment was used to illustrate the use of MSM compared with other commonly used methods. The worst PU Category recorded across all skin sites within a patient and up to the present assessment was recorded at each assessment.

The 1846 patients provided a median (range) of 62 (2, 182) assessments. The median (range) number of days between visits was 4 (1, 42) days and the length of follow-up ranged from 1 to 65 days, with median 14 days, and an interquartile range (7, 25) days. At the patient level, there were 9975 assessments and 8129 transitions were observed as shown in M_1 and M_0 (5), where 1 denotes intervention and 0 denotes the control group. The observed transitions included 45 from state 3 (Category I) to state 4 (Category II+), overall incidence of new Category II PUs was equal to 7.2% (Table 2). There were 11 transitions observed from state 1 to state 3, 11 transitions from state 1 to state 4 and 76 transitions from state 2 to state 4. This is a key feature of panel data whereby the disease process is not fully observed, and transitions between states may occur between assessments.

$$M_1 = \begin{pmatrix} 303 & 67 & 4 & 6 \\ 0 & 2792 & 76 & 30 \\ 0 & 0 & 692 & 17 \\ 0 & 0 & 0 & 0 \end{pmatrix} M_0 = \begin{pmatrix} 347 & 82 & 7 & 5 \\ 0 & 2953 & 80 & 46 \\ 0 & 0 & 594 & 28 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{5}$$

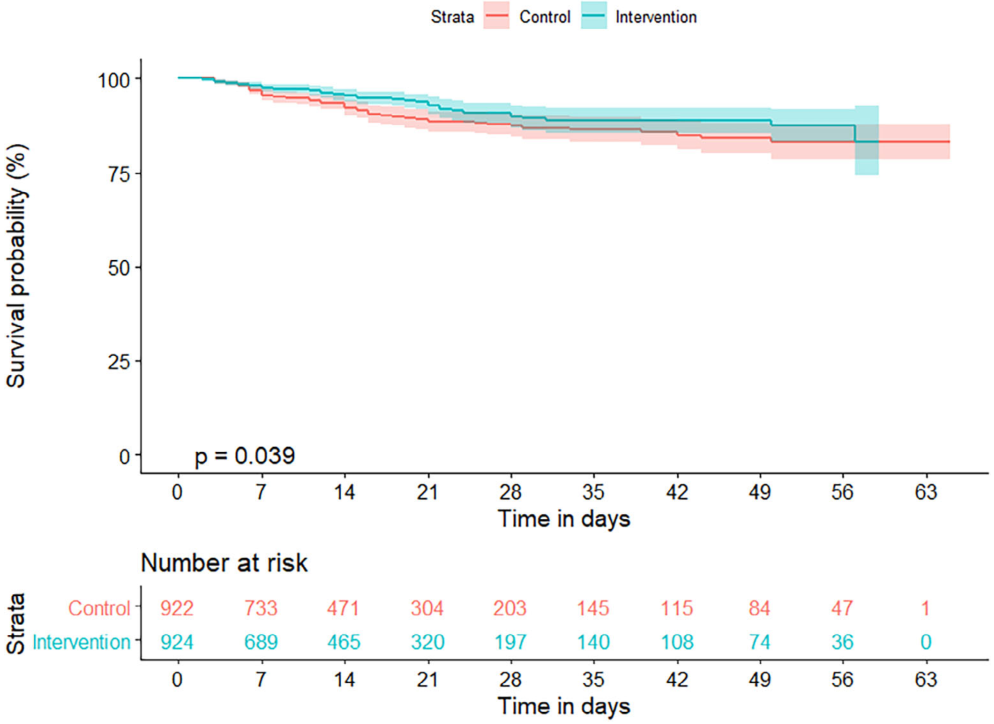
TABLE 2 PRESSURE2 Incidence of new Category II (or more severe) PUs

	Intervention	Control	Total
New Category II+ PU	53 (5.7%)	79 (8.6%)	132 (7.2%)
No new Category II+ PU	871 (94.3%)	843 (91.4%)	1714 (92.8%)
Total	924 (100.0%)	922 (100.0%)	1846 (100.0%)

TABLE 3 Model-based analysis of PRESSURE2

	Point estimate of treatment effect	95% CI
Logistic regression (odds ratio)	0.65	(0.45, 0.93)
Cox PH (hazard ratio)	0.69	(0.49, 0.98)
Multistate model		
β_{12}	0.98	(0.72, 1.32)
β_{23}	0.90	(0.70, 1.15)
β_{34}	0.57	(0.40, 0.81)

FIGURE 2 Kaplan-Meier plot for the time to incidence of a new Category II or more severe PU in the PRESSURE2 trial, based on a subset of patients during their treatment phase [Color figure can be viewed at wileyonlinelibrary.com]



Initial analyses of the PRESSURE2 data focused on incidence of a new Category II+ PU using logistic and Cox PH regression models (Table 3). The estimands in these analyses were the odds ratio for incidence of, and the hazard ratio for time to, a Category II+ PU, respectively. Both estimates were significantly different from one, which suggests that treatment does have an influence on Category II+ PU development. However, the binary outcome does not take into account the time a patient is in the trial before discharge and there is evidence from the Kaplan-Meier curves in Figure 2 that the proportional hazards assumption is not valid. Specifically, the treatment effect does not appear until after the first week at risk; such a delayed effect might be expected in a prevention trial, in which severe ulcers take some time to develop, with a corresponding delay in evidence of prevention.

The 4-state model illustrated in Figure 1 was fitted to the full data set and indicated that the treatment effect is not statistically significant on transitions between healthy and altered skin and between altered skin and Category I PU, but there was a substantial and significant treatment effect for the transition between Category I and Category II+. This finding is consistent with the Kaplan-Meier estimates in Figure 2, as it shows that the treatment effect is mainly on the later transition and is only evident when patients have passed through the intermediate states.

The results from this analysis indicate that there is merit in using the longitudinal data to understand the natural history of the disease and to identify where treatment may have most benefit. There are differences in the estimated

TABLE 4 Factors varied in simulation study

Total sample size	Length of follow-up	Assessment frequency	Baseline transition intensities $\mathbf{q(0)} = (q_{12.0}, q_{23.0}, q_{34.0})$	Treatment effects (hazard ratios) $Exp(\beta) = (e^{\beta_{12}}, e^{\beta_{23}}, e^{\beta_{34}})$
100	60 days[a]	Daily[a]	(0.05,0.05,0.03) [a]	(1, 1, 1)
200	30 days	Every 2 days	(0.01,0.01,0.01)	(0.67, 0.67, 0.67) [a]
500	14 days	Every 3 days	(0.01,0.01,0.05)	(0.5, 0.5, 0.67)
1000	7 days	Every 7 days	(0.01,0.05,0.01)	(0.67, 0.67, 0.5)
2000		Every 14 days	(0.05,0.01,0.01)	(0.9, 0.9, 0.67)
			(0.01,0.05,0.05)	
			(0.05,0.01,0.05)	
			(0.05,0.05,0.01)	

[a] Base case settings.

treatment effect for different transitions, which are obscured by the use of a model with a single outcome, such as time to event. It is of interest to understand how a multistate model, which is able to estimate treatment effects at different stages of the disease process, could be used to inform the design and analysis of a future RCT.

4 | SIMULATION STUDY

A review of published PU prevention trials showed that they have included as few as 10[26] and as many as 2029[24] patients, with a median size of 75 patients. Follow-up times have ranged from a few days[27] to 2 or more months,[24] with the frequency of assessments ranging from daily[28] to once a week[27] or less frequently.[29] This simulation study shows how MSM might be used to optimize power calculations based on these design features. Extensive simulations exploring the impact on the power of a trial using MSM compared with analyses with a single endpoint are summarized. Model outputs from the analysis of PRESSURE2 have been used to inform the design of this simulation study.

4.1 | Design of the simulation study

The aim of the simulation study was to assess the impact on power and sample size of using different statistical models and methods to analyze data collected in PU prevention trials such as that described in the motivating example and illustrated in Figure 1. A number of scenarios, based on varying the sample size, length of follow-up, assessment frequency, baseline transition intensities, and treatment effects, were considered (Table 4).

The base case assumed that, patients were followed up daily for a maximum of 60 days, with a moderate treatment effect ($\beta_{12} = \beta_{23} = \beta_{34} = 0.67$), high risk of transitions 1 → 2 and 2 → 3 and a moderate risk of transition 3 → 4 ($q_{12.0} = q_{23.0} = 0.05$, $q_{34.0} = 0.03$). In all scenarios, the proportions of patients in states 1 (healthy), 2 (preclinical), and 3 (mild disease) at baseline ($t = 0$) were 15%, 70%, and 15%, respectively. In each scenario, patients were equally allocated to one of two treatment groups and the censoring rate was assumed to be 0.05 per unit time for all transitions.

The methods evaluated under each scenario were the binary logistic regression model, Cox PH model, and four MSM whereby the treatment effects were either,

a unconstrained, that is, $\beta_{12} \neq \beta_{23} \neq \beta_{34}$,
b completely constrained, that is, $\beta_{12} = \beta_{23} = \beta_{34}$,
c partially constrained, that is, $\beta_{12} = \beta_{23} \neq \beta_{34}$, or
d partially constrained, that is, $\beta_{12} \neq \beta_{23} = \beta_{34}$.

Treatment effects in the binary logistic regression and Cox PH model were assessed using the Wald statistic and significance concluded at the 5% level. Similarly, for the completely constrained MSM, which has a single common treatment effect, the Wald statistic from the maximum likelihood estimation was calculated. The unconstrained and partially constrained MSM had three and two treatment effects, respectively, so that Hochberg's multiple testing procedure based on Bonferroni corrections was adopted in order to maintain the overall 5% type I error.[30] In this case, empirical power was reported overall by examining the Wald statistic for the treatment effect on each transition, for example, for the unconstrained model 5% significance was concluded if either (i) all three transitions were significant at the 5% level, or (ii) at least two treatment effects were significant at the 2.5% level, or (iii) at least one treatment effect was statistically significant at the 1.67% level. Bias of the estimates and coverage were also calculated in addition to the Monte Carlo standard error in line with recommendations for simulation studies.[31,32]

A total of 1000 simulations were conducted for each scenario. The same data sets were used to compare statistical methods but different data sets were generated for each scenario being considered.

4.2 | Results

For the null case, where data sets were generated assuming no treatment effect, the type I error was close to 5%, as expected, in each multistate model and in the logistic and Cox PH regression models. For the base case, where the treatment effect was 0.67 on each transition, all MSM had greater power compared with the binary logistic regression model and the Cox PH model. For example, with 500 patients the binary and Cox models provide power of 57.5% and 68%, respectively, the multistate model with no constraint on the treatment effect provides power of 72.5% and MSM with some constraint(s) applied to the treatment effect provide a minimum of 80% power in this case (Figure 3).

4.3 | Length of follow-up

The simulation study explored lengths of follow-up of 7, 14, 30, days and 60 days (the base case) with all other parameters remaining as in the base case. In all cases the MSM had greater power than the corresponding Cox and binary logistic regression analyses when applied to data with the same follow-up periods. Figure 4 shows results for the unconstrained multistate model with various durations of follow-up compared with logistic and Cox models with 60 days follow-up. The results indicate that a follow-up period of 60 days provides some additional efficiency compared with a follow-up period of 30 days when using a multistate model, while a follow-up period of 7 or 14 days leads to substantially reduced power, largely due to the low number of transitions to the absorbing state. Notably, the unconstrained multistate model with 30-day follow-up had similar power to a Cox model with data collected for 60 days (Figure 4).

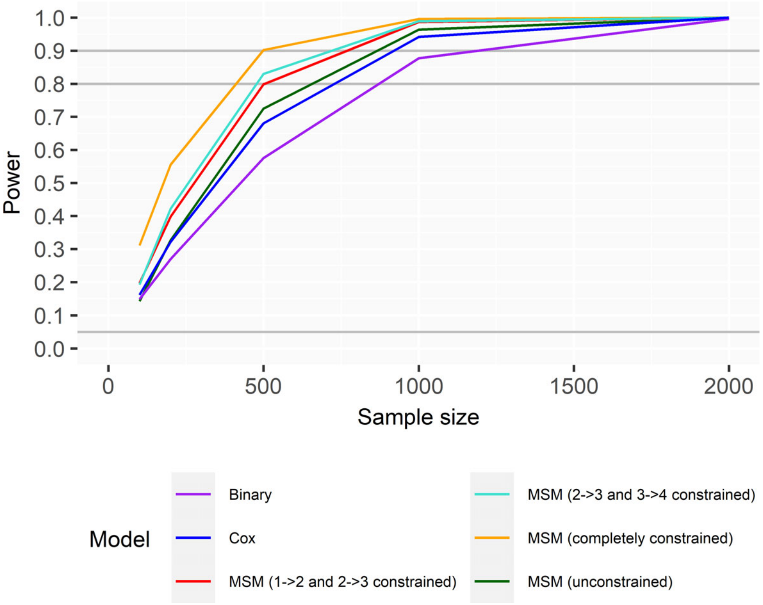


FIGURE 3 Power of detecting a significant treatment effect overall according to sample size for the base case (maximum length of follow-up = 60 days, assessment frequency = daily, $\exp(\boldsymbol{\beta}) = (0.67, 0.67, 0.67)$, $\mathbf{q}(0) = (0.05, 0.05, 0.03)$) [Color figure can be viewed at wileyonlinelibrary.com]

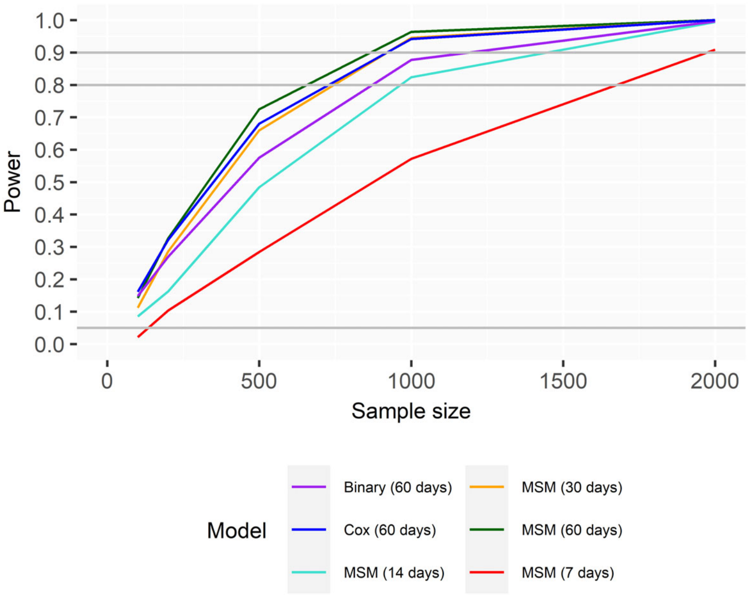


FIGURE 4 Power of detecting a significant treatment effect overall according to sample size for different lengths of follow-up (assessment frequency = daily, $\exp(\beta) = (0.67, 0.67, 0.67)$, $q = (0.05, 0.05, 0.03)$) [Color figure can be viewed at wileyonlinelibrary.com]

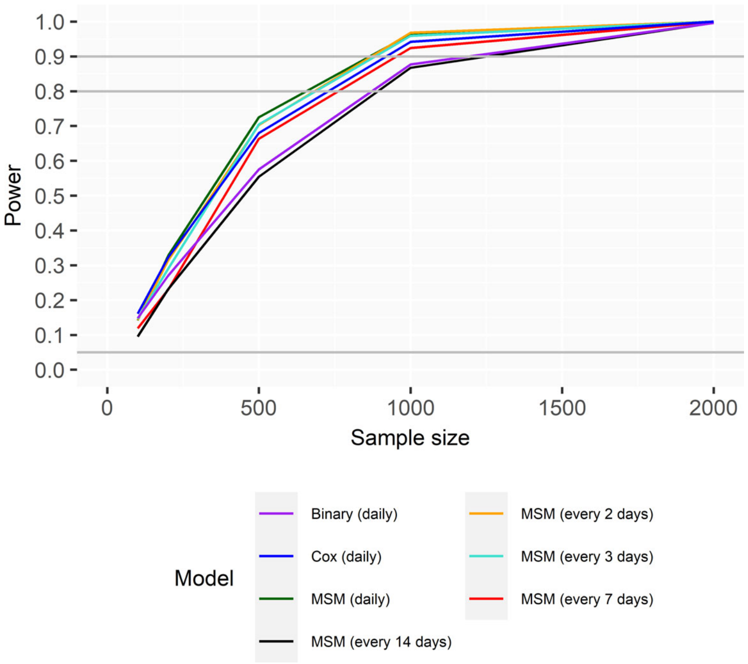


FIGURE 5 Power of detecting a significant treatment effect overall according to sample size for different intervals between assessments (maximum length of follow-up= 60 days, $\exp(\beta) = (0.67, 0.67, 0.67)$, $q = (0.05, 0.05, 0.03)$) [Color figure can be viewed at wileyonlinelibrary.com]

4.4 | Assessment intervals

Assessment intervals of daily, every 2 days, every 3 days, every 7 days, and every 14 days were considered with all other parameters remaining as in the base case including planned follow-up of 60 days. The results indicate that MSM fitted to assessments taken daily, every 2 days or every 3 days, perform at least as well as Cox models applied to data collected daily. There is a large improvement in efficiency from using a multistate model compared with a binary logistic regression model in these scenarios. For example, to achieve 80% power, a multistate model would require around 650 patients with data collected daily or every 2 to 3 days, whereas data would need to be collected daily for an additional 200 (approximately) patients to provide similar levels of power using Binary logistic regression (Figure 5).

4.5 | Different treatment effects and baseline transition intensities

Table 5 summarizes results from scenarios governing treatment effects and transition intensities. MSM conferred substantially increased power compared with the binary logistic and Cox PH regression models when the baseline intensity

TABLE 5 Estimate of power for combinations of baseline transition intensities (maximum length of follow-up = 60 days, assessment frequency = daily, treatment effects [hazard ratios] Exp(β) = (0.67,0.67,0.67))

		Baseline transition intensities			Estimate of power (Monte Carlo SE)		
N	**Scenario**	$q_{12.0}$	$q_{23.0}$	$q_{34.0}$	**Binary logistic regression**	**Cox PH model**	**Overall multistate model**
500	1	0.01	0.01	0.01	0.137 (0.0109)	0.137 (0.0109)	0.177 (0.0121)
	2	0.01	0.01	0.05	0.275 (0.0141)	0.315 (0.0147)	0.270 (0.0140)
	3	0.01	0.05	0.01	0.289 (0.0143)	0.375 (0.0148)	0.597 (0.0155)
	4	0.05	0.01	0.01	0.152 (0.0114)	0.152 (0.0114)	0.265 (0.0140)
	5	0.01	0.05	0.05	0.597 (0.0155)	0.738 (0.0139)	0.716 (0.0143)
	6	0.05	0.01	0.05	0.631(0.0153)	0.762 (0.0135)	0.728 (0.0141)
	7	0.05	0.05	0.01	0.309 (0.0146)	0.339 (0.0150)	0.669 (0.0149)
	8	0.05	0.05	0.05	0.684 (0.0147)	0.824 (0.0120)	0.775 (0.0132)
1000	1	0.01	0.01	0.01	0.301 (0.0145)	0.318 (0.0147)	0.422 (0.0156)
	2	0.01	0.01	0.05	0.550 (0.0157)	0.573 (0.0156)	0.597 (0.0155)
	3	0.01	0.05	0.01	0.559 (0.0157)	0.586 (0.0156)	0.905 (0.0093)
	4	0.05	0.01	0.01	0.304 (0.0145)	0.308 (0.0146)	0.541 (0.0158)
	5	0.01	0.05	0.05	0.903 (0.0094)	0.955 (0.0066)	0.959 (0.0063)
	6	0.05	0.01	0.05	0.900 (0.0095)	0.958 (0.0063)	0.962 (0.0061)
	7	0.05	0.05	0.01	0.576 (0.0156)	0.605 (0.0155)	0.940 (0.0075)
	8	0.05	0.05	0.05	0.916 (0.0088)	0.978 (0.0046)	0.980 (0.0044)
2000	1	0.01	0.01	0.01	0.543 (0.0158)	0.551 (0.0157)	0.779 (0.0131)
	2	0.01	0.01	0.05	0.808 (0.0125)	0.838 (0.0117)	0.915 (0.0088)
	3	0.01	0.05	0.01	0.846 (0.0114)	0.871 (0.0106)	0.997 (0.0017)
	4	0.05	0.01	0.01	0.556 (0.0157)	0.562 (0.0157)	0.873 (0.0105)
	5	0.01	0.05	0.05	0.995 (0.0022)	1.000 (0.0000)	1.000 (0.0000)
	6	0.05	0.01	0.05	0.999 (0.0010)	1.000 (0.0000)	1.000 (0.0000)
	7	0.05	0.05	0.01	0.873 (0.0105)	0.901 (0.0094)	0.999 (0.0010)
	8	0.05	0.05	0.05	0.997 (0.0017)	1.000 (0.0000)	1.000 (0.0000)

for the transition from state 3 to state 4 was low ($q_{34.0} = 0.01$). A consistent increase in power was observed in scenarios 1, 3, 4, and 7, whereas there were similar levels of power observed for each model under scenarios where the baseline transition intensity from state 3 to state 4 was high (Category I to II+, $q_{34.0} = 0.05$). In some cases (eg, N = 500, scenario 2) lower power was observed for the multistate model compared with the binary logistic regression and Cox PH models. We note, for example, that the Cox PH model estimates the treatment effect on the transition from any of the states 1, 2, or 3 to state 4 and significance testing is conducted at the 5% level. In contrast, the multistate model estimates the treatment effect on individual transitions (ie, $1 \rightarrow 2$, $2 \rightarrow 3$, and $3 \rightarrow 4$) and significance testing is conducted according to Hochberg's method for multiple testing. Therefore, the Cox PH model is expected to perform as well as the overall multistate model in situations when the baseline transition intensity to the absorbing state is high and may therefore be the preferred method for primary analysis as it requires less computing power, and is widely understood. Similar results are observed in Appendix S1 for different treatment effects.

5 | DISCUSSION

5.1 | Summary of results

RCTs of strategies for prevention of diseases and medical conditions often involve repeated assessments of the severity of disease at multiple time-points. The potential estimands from different models that could be applied to data of this structure include odds ratios for a binary endpoint, hazard ratios for a time to event endpoint, and transition-specific hazard ratios obtained from MSM. Secondary analysis of the full data from an existing PU trial using a MSM showed that such models can provide a deeper understanding of PU natural history and how treatment acts at each stage of the disease pathway. Where such data are collected they should be analyzed in detail in order to understand the mechanisms of action of different treatments.

Simulations have shown that, depending on the estimand of interest, analysis using MSM can have a substantial impact on power, or equivalently a reduction in sample size. Note that there is no formula for calculating the sample size for a trial using MSM as the primary analysis and simulations such as those used in this article can be used instead.

5.2 | Implications on clinical settings

The results of the simulation study suggest that in PRESSURE2 the length of follow-up could have been reduced to 30 days or assessments conducted every 2 or 3 days to provide similar levels of power as would be obtained by Cox or binary logistic regression models applied to daily measurements for 60 days. This has the potential to reduce trial resource use by using fewer patients, with savings in clinical research nurse time and data management. In many scenarios, fewer patients need to be recruited overall and fewer are unnecessarily exposed to inferior treatments. Moreover, evidence of treatment effectiveness (or not) will emerge more quickly leading to quicker changes in practice for subsequent patients. However, this should be considered in conjunction with the relevant clinical research question (estimand) since the overall significance level for MSM reflects treatment effects across all transitions. For example, if primary interest lies in preventing Category II+ PUs then the commonly used methods may be sufficient, and have the advantage that the resulting significance level is directly related to a single treatment effect. If however interest lies in assessing whether the treatment can reduce transitions at any stage along the pathway, and an overall model significance level is acceptable, then MSM will lead to more efficient designs at lower cost. Furthermore, health economic analyses commonly use Markov models[13] and by using MSM for the analysis of the main trial, the results of the clinical research and health economics research will be better aligned.

Our simulation study provides a comprehensive set of results under a range of scenarios and compared MSM with simpler models (logistic regression and the Cox PH model), in addition to reviewing the impact of applying constraints to treatment effects within MSM. Constraining treatment effects to be equivalent within the MSM did provide additional power but should be used with caution as they may not be a realistic representation if the true treatment effects differ between disease stages. Although we were motivated by trials in PUs, the methods are general and can be applied to any setting whereby a disease process can be reasonably represented by a multistate model.

5.3 | Some statistical considerations

It is important to try to understand how the additional power from MSM arises, given that the number of Category II+ PUs observed was the same for all analyses. It is clear that more data on early skin changes is included in MSM; this, together with the structure of the model which links the different transitions together, is the source of the additional information. If the MSM is not consistent with the observed natural history of the disease of interest, then either the predicted increase in power will not manifest, or spurious increases in power will result. Therefore, it is imperative that a good model is adopted and (in line with good statistical practice) the fit of the model is checked carefully. The Appendix provides plots of the observed and model-fitted prevalence in each of the four states in the PRESSURE2 trial analysis, with reassuring agreement between them. This shows that the Markov assumption (transitions depend only on current disease stage, time since randomization, and covariates) holds over the 60-day duration of study. For disease prevention studies with much longer time horizons, such as cancer prevention studies, this assumption may not hold and alternatives may be required. The msm software in R allows for piecewise continuous Markov models which may be more appropriate but increases the number of parameters to be estimated.

All analyses in the article were conducted in R using freely available, general software, which is important if these methods are to be used more widely. Other statistical software programs contain packages that will fit MSM[10,33] but most focus on the case where exact transition times are known, with no interval censoring, and on semi Markov models, where time is reset to zero when a new state is entered. These cases simplify the likelihood and implementation but are not suitable for the case of PU prevention.

In theory, where trial patients are all followed up at the same time points, discrete time models may be suitable. In these cases the model is specified through the probabilities of moving states between scheduled time points, rather than transition intensities. Resulting likelihoods include multinomial terms and can be fitted using standard software, but results will be specific to the observation times.

5.4 | Future research

This simulation study makes a number of assumptions, including that censoring patterns are independent of skin status or patient condition, and that the MSM allows progression only. In reality, patients may move between states in both directions and this should be considered for further research. Furthermore, this article has studied the use of MSM on a patient level "worst observed skin state" basis, whereas assessments made on all 14 skin sites could be used to provide additional power. MSM have been used for correlated disease processes of this type, for example, psoriatic arthritis,[34] but resulting models are more complex and there is little available software in the case of interval censoring.

Misclassification of (particularly) early stage disease is a major concern in the PU setting because of the subjective nature of skin assessment.[20,35] The extent of misclassification between categories and the impact of this on assessment of treatment effects is the subject of further work.

Missing data methods need to be considered in this setting. Skin classifications that are not recorded may be dependent on the PU stage itself. Failure to explicitly model this missing data mechanism may result in biased estimates of the rate of PU onset and change. The extent of the potential bias needs to be examined to inform PU assessment strategies and analysis of future trials.

In conclusion, this simulation study demonstrated that current methods of analysis of PU prevention trials may be inefficient, requiring large sample sizes and frequent assessments. MSM have the potential to maximize the information collected in RCTs with serial disease category assessments and could transform the design of clinical trials of PU prevention strategies, although further methodological work is required to provide robust recommendations for the design and analysis of such trials in general. Moreover, analysis of the full data can provide important insight into the mechanisms of action of the treatment under evaluation.

ACKNOWLEDGEMENTS

Isabelle L. Smith is funded by a National Institute for Health Research (NIHR) Doctoral Fellowship (NIHR DRF-2016-09-085). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

DATA AVAILABILITY STATEMENT

Reasonable requests for data from the PRESSURE2 trial, and code for the simulation study may be directed to the lead author of the paper.

ORCID

Isabelle L. Smith https://orcid.org/0000-0002-8326-1075

REFERENCES

1. Yusuf S, Pfeffer MA, Swedberg K, et al. Effects of candesartan in patients with chronic heart failure and preserved left-ventricular ejection fraction: the CHARM-preserved trial. *Lancet*. 2003;362(9386):777-781.
2. Brass D, Fouweather T, Stocken DD, et al. An observer-blinded randomized controlled pilot trial comparing localized immersion psoralen-ultraviolet a with localized narrowband ultraviolet B for the treatment of palmar hand eczema. *Br J Dermatol*. 2018;179(1):63-71.
3. Jackson CH, Sharples LD. Hidden Markov models for the onset and progression of bronchiolitis obliterans syndrome in lung transplant recipients. *Stat Med*. 2002;21(1):113-128.
4. Ieva F, Jackson CH, Sharples LD. Multi-state modelling of repeated hospitalisation and death in patients with heart failure: the use of large administrative databases in clinical epidemiology. *Stat Methods Med Res*. 2017;26(3):1350-1372.
5. Lauseker M, Hasford J, Hoffmann VS, et al. A multi-state model approach for prediction in chronic myeloid leukaemia. *Ann Hematol*. 2015;94(6):919-927.
6. Conlon ASC, Taylor JMG, Sargent DJ. Multi-state models for colon cancer recurrence and death with a cured fraction. *Stat Med*. 2014;33(10):1750-1766.
7. O'Keeffe AG, Tom BDM, Farewell VT. A case-study in the clinical epidemiology of psoriatic arthritis: multistate models and causal arguments. *J R Stat Soc C-Appl*. 2011;60:675-699.

8. Smith EMD, Eleuteri A, Goilav B, et al. A Markov multi-state model of lupus nephritis urine biomarker panel dynamics in children: predicting changes in disease activity. *Clin Immunol*. 2019;198:71-78.
9. Marqueen KE, Waingankar N, Sfakianos J, et al. Identifying high surgical risk in muscle-invasive bladder cancer (MIBC) patients undergoing radical cystectomy (RC). *J Clin Oncol*. 2018;36(6):460.
10. Crowther MJ, Lambert PC. Parametric multistate survival models: flexible modelling allowing transition-specific distributions with application to estimating clinically useful measures of effect differences. *Stat Med*. 2017;36(29):4719-4742.
11. de Wreede LC, Fiocco M, Putter H. Mstate: an R package for the analysis of competing risks and multi-state models. *J Stat Softw*. 2011;38(7):1-30.
12. Jackson CH. Multi-state models for panel data: the msm package for R. *J Stat Softw*. 2011;38(8):1-28.
13. Price MJ, Welton NJ, Ades AE. Parameterization of treatment effects for meta-analysis in multi-state Markov models. *Stat Med*. 2011;30(2):140-151.
14. Manzini G, Ettrich TJ, Kremer M, et al. Advantages of a multi-state approach in surgical research: how intermediate events and risk factor profile affect the prognosis of a patient with locally advanced rectal cancer. *BMC Med Res Methodol*. 2018;18:23.
15. Cassarly C, Martin RH, Chimowitz M, Pena EA, Ramakrishnan V, Palesch YY. Assessing type I error and power of multistate Markov models for panel data-a simulation study. *Commun Stat Simul Comput*. 2017;46(9):7040-7061.
16. Viúdez A, Carmona-Bayonas A, Gallego J, et al. Optimal duration of first-line chemotherapy for advanced gastric cancer: data from the AGAMENON registry. *Clin Transl Oncol*. 2020;22(5):734-750.
17. Le-Rademacher JG, Peterson RA, Therneau TM, Sanford BL, Stone RM, Mandrekar SJ. Application of multi-state models in cancer clinical trials. *Clin Trials*. 2018;15(5):489-498.
18. Zeng LL, Cook RJ, Lee KA. Design of cancer trials based on progression-free survival with intermittent assessment. *Stat Med*. 2018;37(12):1947-1959.
19. Gruger J, Kay R, Schumacher M. The validity of inferences based on incomplete observations in disease state models. *Biometrics*. 1991;47(2):595-605.
20. Beeckman D, Schoonhoven L, Fletcher J, et al. EPUAP classification system for pressure ulcers: European reliability study. *J Adv Nurs*. 2007;60(6):682-691.
21. Haesler E, ed. National pressure ulcer advisory panel epuap, and pan pacific pressure injury alliance. *Prevention and Treatment of Pressure Ulcers: Clinical Practice Guideline*. Osborne Park, Western Australia: Cambridge Media; 2014.
22. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Stat Med*. 2007;26(11):2389-2430.
23. Ardo van den H. *Multi-State Survival Models for Interval Censored Data*. Boca Raton, FL: CRC Press; 2017.
24. Smith I. Comparing alternating pressure mattresses and high-specification foam mattresses to prevent pressure ulcers in high-risk patients: the PRESSURE 2 RCT. *Health Technol Assess*. 2019;23(52):1-176.
25. McInnes E, Jammali-Blasi A, Bell-Syer SEM, Dumville JC, Middleton V, Cullum N. Support surfaces for pressure ulcer prevention. *Cochrane Database Syst Rev*. 2015;9:CD001735.
26. Motta G, Dunham L, Dye T, Mentz J, O'Connell-Gifford E, Smith E. Clinical efficacy and cost-effectiveness of a new synthetic polymer sheet wound dressing. *Ostomy Wound Manage*. 1999;45(10):41-44.
27. Cooper PJ, Gray DG, Mollison J. A randomised controlled trial of two pressure-reducing surfaces. *J Wound Care*. 1998;7(8):374-376.
28. Demarre L, Beeckman D, Vanderwee K, Defloor T, Grypdonck M, Verhaeghe S. Multi-stage versus single-stage inflation and deflation cycle for alternating low pressure air mattresses to prevent pressure ulcers in hospitalised patients: a randomised-controlled clinical trial. *Int J Nurs Stud*. 2012;49(4):416-426.
29. Ford CN, Reinhard ER, Yeh D, et al. Interim analysis of a prospective, randomized trial of vacuum-assisted closure versus the healthpoint system in the management of pressure ulcers. *Ann Plas Surg*. 2002;49(1):55-61.
30. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*. 1988;75(4):800-802.
31. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate. *Stat Methods*. 2019;38(11):2074-2102.
32. Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Stat Med*. 2006;25(24):4279-4292.
33. Wu HM, Yen MF, Chen THH. SAS macro program for non-homogeneous Markov process in modeling multi-state disease progression. *Comput Meth Prog Bio*. 2004;75(2):95-105.
34. O'Keeffe AG, Tom BDM, Farewell VT. Mixture distributions in multi-state modelling: some considerations in a study of psoriatic arthritis. *Stat Med*. 2013;32(4):600-619.
35. Nixon J, Thorpe H, Barrow H, et al. Reliability of pressure ulcer classification and diagnosis. *J Adv Nurs*. 2005;50(6):613-623.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Smith IL, Nixon JE, Sharples L. Power and sample size for multistate model analysis of longitudinal discrete outcomes in disease prevention trials. *Statistics in Medicine*. 2021;1–12. https://doi.org/10.1002/sim.8882