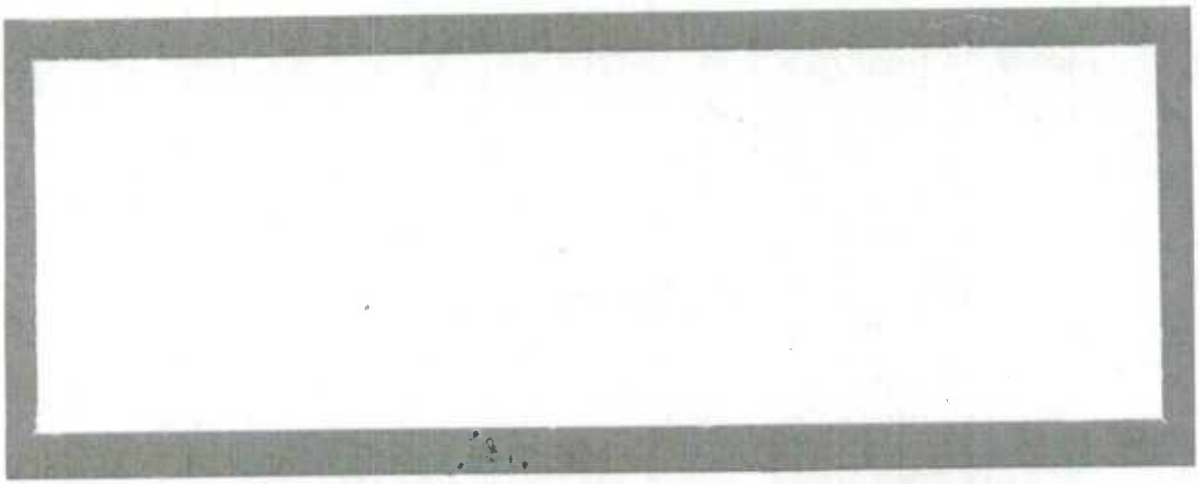
Statistics Canada Statistique Canada



Methodology Branch

Social Survey
Methods Division

Direction de la méthodologie

Division des méthodes
d'enquêtes sociales

Canada

WORKING PAPER
METHODOLOGY BRANCH

A THEORY OF LONGITUDINAL MICROSIMULATION AND ITS APPLICATION TO OPTIMAL POLICY COMPARISONS

SSMD - 94 - 005 E

Gurupdesh S. Pandher and Milorad S. Kovacevic

Methods Development and Analysis Section
Social Survey Methods Division, Methodology Branch
Statistics Canada, Ottawa, K1A 0T6

August 1994

(Version II)

The views presented in this paper are those of the authors and do not necessarily represent the views or policies of Statistics Canada.

A THEORY OF LONGITUDINAL MICROSIMULATION AND ITS APPLICATION TO OPTIMAL POLICY COMPARISONS

G.S. Pandher and M.S. Kovacevic
Methods Development and Analysis Section
Social Survey Methods Division, Methodology Branch
Statistics Canada, Ottawa, K1A 0T6
Canada

Abstract

With advances in computing power, longitudinal microsimulation modelling (MSM) has become an important analytical and inferential device for quantitative policy analysis in a wide range of policy development and decision-making settings. It allows a form of controlled experimentation which is now widely applied in disciplines where experiments on real populations are infeasible or sometimes impossible. In this context, microsimulation output has analytical and inferential uses rather than merely providing data to construct simple point estimates for descriptive purposes.

In this paper, first a general theory of the longitudinal microsimulation process is developed. This abstraction is used to identify an optimal microsimulation design permitting the use of statistically efficient policy comparison methodologies. Next, a statistical model for MSM is obtained from the theoretical representation to estimate and test policy effects and to demonstrate the gains from using the new design on important aspects of the statistical testing methodology (eg. power, sample size, significance level). One important outcome of the proposed MSM design is that dramatic reductions in sample size are uniformly achievable over conventional approaches at any specified level of significance and power for the testing procedure.

Results are presented from a simulation study in which the theory and concepts developed in the paper are applied to a small microsimulation model. Remarkable gains in efficiency for the estimator of policy effects were observed under the new microsimulation design.

Keywords: Across-policy Monte-Carlo effects; Decision functions, states, sequences and spaces; Designed experiments; Sample size reduction; Variance reduction.

UNE THÉORIE DE MICROSIMULATION LONGITUDINALE ET SON UTILISATION POUR OPTIMAL COMPARER DES POLITIQUES

G.S. Pandher et M.S. Kovacevic
Division des méthodes d'enquêtes sociales
Direction de la méthodologie
Statistique Canada, Ottawa, K1A 0T6

Résumé

En tant qu'outil servant à élaborer des politiques, la modélisation par microsimulation longitudinale (MML) est un mécanisme d'analyse et d'inférence majeur pour étudier l'impact de diverses politiques.

Dans cet article, on développe une théorie du procédé de microsimulation longitudinale et on l'utilise pour identifier certains aspects du plan qui altèrent le rendement d'estimateurs mesurant la différence entre des politiques. On propose une conception d'un modèle de microsimulation qui permet d'utiliser une méthodologie de comparaison très efficace et sensible. Cette méthodologie sert à évaluer l'impact d'interventions en matière de politique en contrôlant l'interaction entre la variabilité de type Monte-Carlo et la politique retenue. Ensuite on développe un modèle statistique fondé sur la théorie de la microsimulation longitudinale s'appliquant aux caractéristiques des individus. Ce modèle sert à estimer et à tester l'effet de politiques. On démontre les bénéfices résultant de la reformulation d'aspects importants de la méthodologie utilisée pour les tests statistiques.

Enfin, on présente les résultats d'une étude de simulation dans laquelle les concepts présentés dans l'article sont appliqués à un modèle de microsimulation simple. L'étude souligne que le nouveau plan de microsimulation a produit des gains remarquables quant à l'efficacité de l'estimateur des effets de politique.

Mots-clé :
fonctions de décision, états et espaces; séquences de décisions fixes et permutables; Monte-Carlo; plan de génération de biographies; gains en efficacité; test t en paires

TABLE OF CONTENTS

A THEORY OF LONGITUDINAL MICROSIMULATION AND ITS APPLICATION TO OPTIMAL POLICY COMPARISONS

G.S. Pandher and M.S. Kovacevic*

1. INTRODUCTION

With recent advances in computing power, longitudinal microsimulation modelling (MSM) is becoming an increasingly popular, and in many cases indispensable, tool for quantitative policy analysis in a wide range of policy development and decision-making arenas. It allows a form of controlled experimentation which is now widely applied in disciplines where experiments on real populations are infeasible or sometimes impossible. In this context, microsimulation output has analytical and inferential uses rather than merely providing data to construct simple point estimates for descriptive purposes.

MSM affords the researcher the ability to model the essential determinants of complex real world phenomena under various assumptions under his control. An abundant collection of papers with applications of longitudinal microsimulation to study various processes in the economics, finance, demography, public health, energy planning and distributing, etc. are given in Orcutt *et al.* (1986) and Pawlikowski (1990).

Our work originally arose in the context of the Population Health Microsimulation (POHEM) system (Wolfson and Berthelot, 1992), a microsimulation modelling system developed at Statistics Canada which enables the analyst to simulate the Canadian population's health, cost, and medical resource utilization under various policy scenarios. However, we soon realized that a large gap existed in a formalized and unified treatment of longitudinal microsimulation; with most approaches usually treating the subject as an ad-hoc exercise in empirical data analysis.

This paper attempts to fill this void by developing an abstraction for the MSM process and further uses it to identify an optimal microsimulation design permitting the use of statistically efficient policy

---

*Gurupdesh Pandher is Methodologist and Milorad Kovacevic is Senior Methodologist, Social Survey Methods Division, Methodology Branch, Statistics Canada, 16th Floor, Robert Coates Building, Ottawa, Ontario K1A OT6, Canada. The authors would like to acknowledge the support of Social Economic Studies Division in funding this work. We also thank Geoff Rowe, Harold Mantel, and Jean-Marie Berthelot for their beneficial comments in the development of this paper.

comparison methodologies. MSM output generated using the proposed microsimulation design allows a highly efficient and powerful statistical analysis to be performed in assessing the impact of different policy scenarios. One important result obtained is that, under the proposed design, dramatic reductions in sample size may be achieved over conventional approaches at any fixed level of significance and power for the testing procedure.

MSM generates a sample of simulated individual biographies over time. An individual biography may be viewed as a series of simulated event outcomes mimicking the essential features of a complex stochastic process. The only source of randomness or variability induced in this process is due to the randomness of the pseudo-random numbers drawn to execute the various decisions/events in the biography's life-path. We refer to this as Monte-Carlo variability. Moreover, each MSM run is performed under a set of external (policy) conditions under the control of the researcher. Frequently, one is interested in studying the impact of certain changes in the policy environment on the simulated population. Hence, the issue of policy comparisons arises in a context where at least two simulation runs - each under a different policy setting - are performed[2].

The execution of more than one simulation run introduces additional across-policy Monte-Carlo variability in the estimators used to test policy differences. This leads to the confounding of policy effects with across-policy Monte-Carlo variability, diluting the efficiency of estimators and ability of statistical tests to discern the real impact of policy changes - whether observed differences in response are due to policy changes, or are they the result of uncontrolled differences in the random numbers used.

In the next sections we develop a general theoretical representation for longitudinal microsimulation and use it to identify design issues which lower the efficiency of estimators of policy differences. We then propose a biography creation strategy which reduces across-policy Monte-Carlo contamination and allows sharper discernments of policy effects to be made. Based on the theory, a statistical model for MSM is developed to permit estimation and comparison of policy effects and to demonstrate the efficiency gains of the new design. Finally, results from a simulation study are presented in which the theory and concepts developed in the paper were applied to a small microsimulation model.

---

[2] For a formal investigation into the statistical properties (bias, efficiency, stability) of methods aimed at reducing Monte-Carlo variability in estimators based on outcomes from a single simulation run see Kovacevic and Pandher (1993); Schmeiser (1982) studies the effects of batch size on analysis of simulation output.

2. GENERAL REPRESENTATION OF LONGITUDINAL MSM

A creation of a synthetic biography generated at time t may be depicted using the concepts of 'state' and 'laws of motion' found in physics as proposed by Wolfson (1992). The state in the MSM context describes for each individual certain variables of interest (such as age, risk factors, health, etc.) associated with the individual as well as relationships among individuals (such as spouse). The laws of motion specify rules of how the state of each individual may change over time. The velocity with which these laws operate may change over time as the attributes describing the individual's state change. For instance, probabilities of certain events such as death due to coronary heart disease (CHD) may change as risk factors linked to CHD vary over time. The MSM takes on the character of a computer algorithm which embodies these laws of motions in the form of structural connections between temporal states.

In order to proceed further in specifying a more precise mathematical formulation of MSM, we need to further extend the "laws of motion" analogy and distinguish more clearly between two classes of states. State attributes which collectively describe an individual may be separated into two categories. The first category, symbolized by the vector $\underset{\sim}{\alpha}_t = (\alpha_t^{(1)}, \ldots, \alpha_t^{(K_t)})$, constitutes a vector of descriptors and enters the MSM as an exogenous information set at time t . Values in $\underset{\sim}{\alpha}_t$ are not modified by the MSM model. Examples of these variables are disease risk factors, socio-economic status, genetic predisposition, and survival distributions. Note that this exogenous information is usually obtained/estimated from observational and controlled studies. This information is available prior to microsimulation and enters the MS model as an external assumption.

The second category of state variables are termed as endogenous state variables. They describe the state or condition of each individual i generated by MSM and are effected by the exogenous state vector $\underset{\sim}{\alpha}_t$. These endogenous state variables are symbolized by the vector $\underset{\sim}{\beta}_{it} = (\beta_{it}^{(1)}, \ldots, \beta_{it}^{(S_t)})$ and are observed only once the t^{th} time period has been simulated. Components of $\underset{\sim}{\beta}_{it}$ may be variables describing the disease and health status of individual i , utilization of medical resources, medical costs, employment status, income earned, etc. at time t .

Decomposing the total state vector for each biography i at time t in terms of exogenous and endogenous attributes allows the simulated outcome at time t to be portrayed by the macro-level representation given in Figure 1.

Figure 1: Inputs and Outputs of MSM for Biography i at time t.

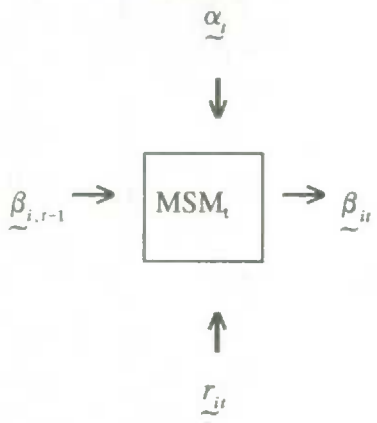


Figure 1 depicts that at time t, the MSM algorithm takes as inputs the parameters consisting of the exogenous state vector $\underset{\sim}{\alpha}_t$ and the endogenous state vector observed at the previous time period $\underset{\sim}{\beta}_{i,t-1}$ and, lastly, a vector of random uniform variates $\underset{\sim}{r}_{it}$ (discussed below). It produces as output (using the laws of motion defined in the MSM) the endogenous outcome states $\underset{\sim}{\beta}_{it}$ at time t. In the microsimulation of biography i at time t, the outcome state $\underset{\sim}{\beta}_{it}$ depends not only on the exogenous conditions prevailing at time t, but also on the outcome state $\underset{\sim}{\beta}_{i,t-1}$ occupied by the biography at the last simulated time point[3].

When performing the microsimulation of a biography at time t, a sequence of uniform random variates $\underset{\sim}{r}_{it}$ (of size m_{it}) is drawn to determine the transition $\underset{\sim}{\beta}_{i,t-1}$ to $\underset{\sim}{\beta}_{it}$.

In order to complete the general description, the formulation of MSM output from the generated outcome states $\underset{\sim}{\beta}_{it}$, $t=1, \ldots, T_i$, needs to be discussed. Time-aggregated output of interest for each biography i is denoted by the vector $\underset{\sim}{y}_i = (y_{i\cdot}^{(1)}, \ldots y_{i\cdot}^{(K)})$ whose components may be the total survival time, number of visits to physicians, medical costs, etc. The time-aggregated output vector for each individual will be based on the endogenous states $\underset{\sim}{\beta}_i = (\underset{\sim}{\beta}_{i1}, \ldots, \underset{\sim}{\beta}_{iT_i})$ realized for the simulated biography till the time of termination T_i. In fact, $\underset{\sim}{y}_i$ is a deterministic function of $\underset{\sim}{\beta}_{it}$, $t=1, \ldots T_i$:

---

[3] The transition to state $\underset{\sim}{\beta}_{it}$ may depend not only on the previous state $\underset{\sim}{\beta}_{i,t-1}$, but also on the states occupied at the previous times $t-1$, $t-2$, and so on. To spare notation, we take it to mean that $\underset{\sim}{\beta}_{it}$ does not just depend on its endogenous state at the previous time point $\underset{\sim}{\beta}_{i,t-1}$, but also on states observed further back in the simulated history of the individual. Similarly, $\underset{\sim}{\alpha}_t$ may be interpreted as not merely representing the exogenous information set used as input to MS at time t, but includes the set of exogenous descriptors up to time t.

$$\underset{\sim}{y}_i = \sum_{t=1}^{T_i} f(\underset{\sim}{\beta}_{it}). \tag{2.2}$$

For example, if $\beta_{it}^{(k)}$ is defined as the endogenous state indicator variable signifying whether or not the biography entered the death at time t,

$$\beta_{it}^{(k)} \equiv \begin{cases} 1 & \text{if employed in period } (t-1, t] \\ 0 & \text{otherwise} \end{cases}$$

then, $y_{i\cdot}^{(k)} = \sum_{t=1}^{T_i} \beta_{it}^{(k)}$ gives the total number of simulation time units in which biography i^{th} was employed.

Therefore, once the outcome states $\underset{\sim}{\beta}_i = (\beta_{i1}, \ldots \underset{\sim}{\beta}_{iT_i})$ are realized in the microsimulation of the i^{th} biography, the corresponding output of interest $\underset{\sim}{y}_i$ may be directly computed functionally in terms of the time-wise elements of $\underset{\sim}{\beta}_i$. Population parameters may then be estimated by appropriate aggregation over individual values given by $\underset{\sim}{y}_i$.

3. A MICRO-LEVEL ABSTRACTION OF LONGITUDINAL MICROSIMULATION

In the previous section a macro (black box) representation of the longitudinal microsimulation process at time t was developed for the output vector $\underset{\sim}{\beta}_{it} = (\beta_{it}^{(1)}, \ldots, \beta_{it}^{(S_t)})$ as a whole. In this section a micro-level theory of how individual outcomes $\beta_{it}^{(k)}$, $k=1, \ldots, S_t$ at each moment t are generated by MSM is developed. Aside from providing an abstraction of the longitudinal microsimulation process, this description is essential for the results obtained later in the paper.

3.1 Decision Steps and Sequences in Longitudinal Microsimulation

Microsimulation for a given biography at each time t (or over period $(t-1, t]$) involves executing a number of ordered decision steps (functions). Let $U_t = \{\mathrm{D}_{t1}, \mathrm{D}_{t2}, \ldots, \mathrm{D}_{tS_t}\}$ represent the universe of all possible decision steps which may be encountered by a biography during its microsimulation at time t. Decision steps answer questions such as "does the biography smoke at time t", "quantity smoked at t", "income earned at time t", "does the biography survive the simulation period $(t-1, t]$", etc. We further denote by $\underset{\sim}{U}_t = \{\underset{\sim}{D}_{t1}, \underset{\sim}{D}_{t2}, \ldots, \underset{\sim}{D}_{tQ_t}\}$ the set of all possible orderings of decision steps in $U_t = \{\mathrm{D}_{t1}, \mathrm{D}_{t2}, \ldots, \mathrm{D}_{tS_t}\}$. The decision sequences (paths) $\underset{\sim}{D}_{tq}$, $q=1, \ldots, Q_t$, which may arise in the state space are dictated by the "laws of motions" of MSM specifying the interconnections between states at each time node. Each decision sequence $\underset{\sim}{D}_{tq}$ in the set $\underset{\sim}{U}_t$ is constructed as an ordered combination of decision steps from $U_t = \{\mathrm{D}_{t1}, \mathrm{D}_{t2}, \ldots, \mathrm{D}_{tS_t}\}$.

Elements in the endogenous state vector $\underset{\sim}{\beta}_{it} = (\beta_{it}^{(1)}, \beta_{it}^{(2)}, \ldots, \beta_{it}^{(S_t)})$ generated for individual i at

simulation time t are outcomes of a decision sequence $\underset{\sim}{D}_{tq}$ chosen from the finite universe of all possible decision sequences $\underset{\sim}{U}_t = \{\underset{\sim}{D}_{t1}, \underset{\sim}{D}_{t2}, \ldots, \underset{\sim}{D}_{tQ_t}\}$. Having established the general notion of decision steps and decision sequences, these concepts will now be formalized and developed in greater detail.

Definition 3.1: Set of All Possible Decision Sequences $\underset{\sim}{U}$

A decision sequence $\underset{\sim}{D}_{tq}$ is a possible ordering of the decision set $U_t = \{D_{t1}, \ldots, D_{tn_t}\}$ permitted by the "laws of motion" of the MSM at time t. Consequently, the set of all possible decision sequences at time t is denoted by $\underset{\sim}{U}_t = \{\underset{\sim}{D}_{t1}, \underset{\sim}{D}_{t2}, \ldots, \underset{\sim}{D}_{tQ_t}\}$, where Q_t is the number of orderings possible.

Each decision step D_{tk} at time t maps a uniform random variate u, drawn to decide the outcome of D_{tk}, onto a point in its decision space X_{tk} representing the set of all possible decision outcomes (either continuous or discrete). Each element of X_{tk} is in the support of a corresponding probability distribution $f_{D_{tk}/D_{tj}}(\cdot)$, $j<k$, (cdf $F_{D_{tk}/D_{tj}}(\cdot)$) which quantifies the probabilities of transition from decision states in X_{tj}, which we call the source space, to states in the decision space X_{tk}, which we call the destination space. These thoughts describing the k^{th} decision step at time t are expressed more precisely by the definition below.

Definition 3.2: Decision Function (Step) D_{tk}

Given a state in the source space X_{tj} and the transition cdf parameter $F_{D_{tk}/D_{tj}}(\cdot)$, the process by which a decision outcome $x \in X_{tk}$ is generated from a uniform random variate u is defined by the following mapping:

$$D_{tk}(u\,;F_{D_{tk}/D_{tj}}): \quad u \rightarrow x, \quad \textit{where } u \in U[0,1], \quad x \in X_{tk}$$

$$\text{such that } x = F^{-1}_{D_{tk}/D_{tj}}(u) \text{ , if } X_{tk} \text{ continous}$$

$$x = \inf\{y \in X_{tk} \mid F(y) \geq u\} \text{ , if } X_{tk} \text{ discrete.}$$

In the case of a discrete destination space, Definition 3.2 assumes an ordinality to exist among the states in X_{tk}. If a natural ranking of discrete states does not exist, a reasonable ordering may be artificially imposed as long as the transition cdf $F_{D_{tk}/D_{tj}}(\cdot)$ over X_{tk} is well defined and used consistently.

3.2 Types of Decision Sequences

We now address the structure of the decision sequences $\underset{\sim}{D}_{tq}$, $q=1, \ldots, Q_t$ in $\underset{\sim}{U}_t$. Decision steps in the decision sequence $\underset{\sim}{D}_{tq}$ may depend on each other in one of three possible ways discussed below.

3.2.1 Fixed Decision Sequence (FDS)

This is the most straightforward and convenient form of decision step dependency. This situation arises when connections among states in the state space are such that while advancing a biography through time t only one ordering of the decision steps in $D_t = \{D_{t1}, \ldots, D_{tS_t}\}$ is possible. FDS arises when states in each source space are connected to states in only one other destination space. In this case, since $Q_t = 1$, the universe of all possible decision sequence configurations contains only one sequence: $\underset{\sim}{U}_t^{(FDS)} = \{\underset{\sim}{D}_{t1}\}$. Furthermore, the fixed decision sequence at time t may be written as $\underset{\sim}{D}_{t1} = \{D_{t1}, \ldots, D_{tS_t}\}$ if we assume without loss of generality that the decision set D_t for time t is ordered sequentially by the order of decisions taken.

In a MSM model exhibiting a FDS structure in the state space at time t, regardless of what outcome occurs upon taking the first decision step D_{t1}, the next decision step will always be D_{t2}. Similarly, regardless of the outcome from decision step D_{t2}, the next decision step will always be D_{t3}, and so on for subsequent decision steps. Previous decision outcomes may, however, effect future decisions quantitatively as discussed later.

Conditions which must be met for a FDS structure to hold among discrete decision spaces at time t are stated in the following theorem.

Theorem 3.1: Sufficient and Necessary Condition for FDS at time t

For each source decision space X_{tk}, $k = 1, \ldots, S_{t-1}$, there exists a connected destination decision space $X_{t,k+p}$, $p \geq 1$, such that

$$\frac{1}{n_{tk}} \sum_{x_{kq} \in X_{tk}} \sum_{x_{k+p,m} \in X_{t,k+p}} f_{D_{t,k+p} / D_{tk} = x_{kq}} (x_{k+p,m}) = 1$$

where $x_{kq}, q = 1, \ldots, n_{tk}$, and $x_{k+p,m}, m = 1, \ldots, n_{t,k+p}$, are the states in X_{tk} and $X_{t,k+p}$, respectively. In refering to the size of the source decision space n_{tk}, we exclude terminal states in X_{tk} while counting n_{tk}.

For all source decision spaces X_{tk}, $k=1, \ldots, S_t-1$, the equation above checks to see whether transitions from all states in X_{tk} (excepting terminal states) occur to the same, one and only, connected decision space $X_{t,k+p}$, $p > 0$.

Proof

Assume that there are a total of n_{tk} states denoted by $x_{km}, m = 1, \ldots, n_{tk}$ in X_{tk}. Take the non-terminal state $x_k \in X_{tk}$. Since x_k is a non-terminal state and, moreover, the state-space allows transition from a source state x_{km} in X_{tk} to states in only one decision space $X_{t,k+p}$, $p > 0$, the sum of all transitions from x_{k1} in X_{tk} to possible destination states $x_{k+p,m}$, $m = 1, \ldots, n_{t,k+p}$, in $X_{t,k+p}$ must add to one:

$$\sum_{x_{k+p,m} \in X_{t,k+p}} f_{D_{t,k+p}/D_{tk}=x_{k1}}(x_{k+p,m}) = 1 .$$

Now a FDS structure among the decision spaces $\{X_{t1}, \ldots, X_{tS_t}\}$ arises when only one possible ordering of decision steps in $U_t = \{D_{t1}, \ldots, D_{tS_t}\}$ is possible given by $\underset{\sim}{U}_t = \{(D_{t1}, \ldots, D_{tS_t})\}$. This implies that all non-terminal states in X_{t1} are connected to states in X_{t2}, all states in X_{t2} are connected to states in X_{t3} and so on.

Repeating the same step given above for x_{k1} in X_{tk} for all non-terminal states $x_{kq}, q = 1, \ldots, n_{tk}$, in X_{tk}, then yields

$$\sum_{x_{kq} \in X_{tk}} \sum_{x_{k+p,m} \in X_{t,k+p}} f_{D_{t,k+p}/D_{tk}=x_{qk}}(x_{k+p,m}) = n_{tk} .$$

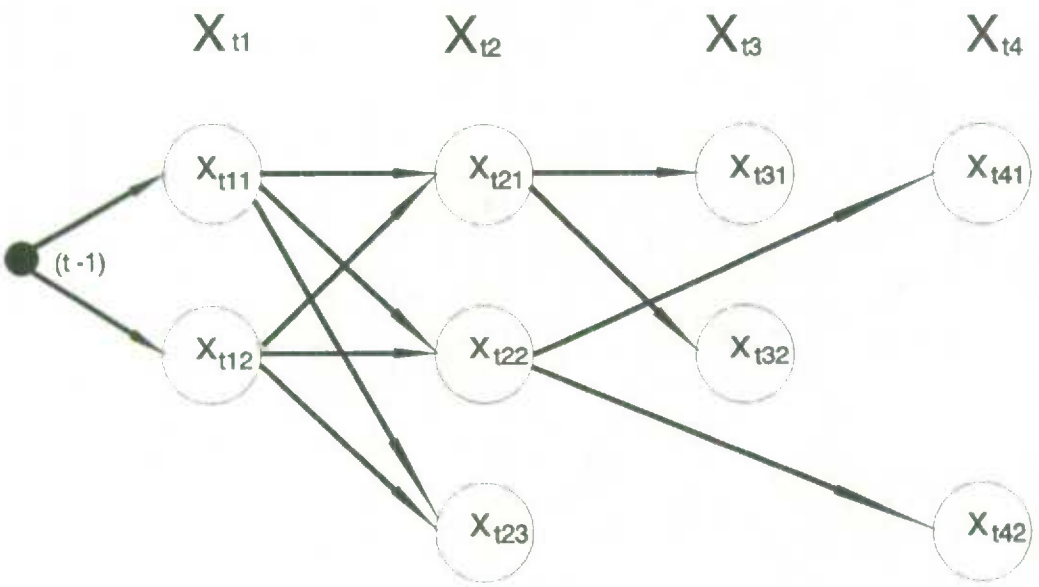
Finally, for the FDS structure to hold across the complete state space at time t, the above expression must hold for all source decision spaces $X_{tk}, k = 1, \ldots, S_t - 1$.

3.2.2 Permutable Decision Sequences (PDS)

Random sequential dependencies may arise in the decision sequences $\underset{\sim}{D}_{tq}, q = 1, \ldots, Q_t$ if the outcome of previous decision steps D_{tk} alters the types of future decisions $D_{t(k+p)}, p \geq 1$ which may be undertaken. This occassion arises if the state space structure is such that a decision space X_{tk} is connected to more than one other decision space. When states in a decision space X_{tk} map onto states in more than one other connected decision space $X_{t,k+p}$, $p \geq 1$, then it is possible for more than one decision path configuration $\underset{\sim}{D}_{tq}, q = 1, \ldots, Q_t, (Q_t > 1)$ to arise. The set of all possible decision sequences possible at time t denoted by $\underset{\sim}{D}_t = \{\underset{\sim}{D}_{t1}, \ldots, \underset{\sim}{D}_{tQ_t}\}$ will now contain more than one element.

As an illustration, suppose that at time t the decision spaces X_{t1}, X_{t2}, X_{t3} and X_{t4} are connected with each other as shown in Figure 2.

Figure 2. Notational Example of Permutable Decision Space Structure at time t.

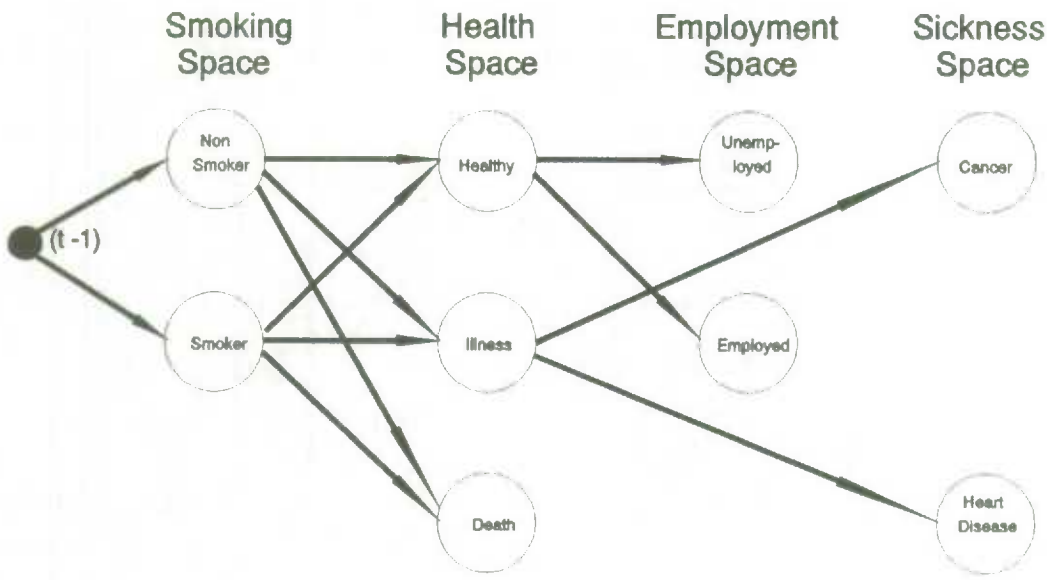


This structure of the decision state space at time t allows the occurence of three decision sequence configurations, namely $\underset{\sim}{D}_{t1} = (D_{t1}, D_{t2})$, $\underset{\sim}{D}_{t2} = (D_{t1}, D_{t2}, D_{t3})$ and $\underset{\sim}{D}_{t3} = (D_{t1}, D_{t2}, D_{t4})$. The set of all possible decision sequences at time t may then be written as

$$\underset{\sim}{U}_t = \{\underset{\sim}{D}_{t1}, \underset{\sim}{D}_{t2}, \underset{\sim}{D}_{t3}\} = \{(D_{t1}, D_{t2}),$$
$$(D_{t1}, D_{t2}, D_{t3}),$$
$$(D_{t1}, D_{t2}, D_{t4})\}$$

The same illustration given above is re-expressed below in terms of a simple (but fictitious) example of a MSM decision space at time t.

Figure 2. Example of Permutable Decision Space Structure at time t.



Conditions which imply a PDS structure among discrete decision spaces at time t are stated in the following theorem.

Theorem 3.2: Sufficient and Necessary Condition for PDS at time t

For at least one source decision space X_{tk}, $k = 1, \ldots, S_{t-1}$, there exists a connected destination decision space $X_{t,k+p}$, $p \geq 1$, such that

$$0 < \frac{1}{n_{tk}} \sum_{x_{kq} \in X_{tk}} \sum_{x_{k+p,m} \in X_{t,k+p}} f_{D_{t,k+p}/D_{tk}=x_{kq}}(x_{k+p,m}) < 1$$

where $x_{kq}, q = 1, \ldots, n_{tk}$, and $x_{k+p,m}, m = 1, \ldots, n_{t,k+p}$ are the states in X_{tk} and $X_{t(k+p)}$, respectively. In refering to the size of the source decision space n_{tk}, we exclude all terminal states in X_{tk} when counting n_{tk}.

The proof of Theorem 3.2 is analogous to that for Theorem 3.1 and may be easily modified for the PDS case. These two theorems may be used algorithmically to distinguish between fixed or permutable decision sequences.

A complication introduced by Permutable Decision Sequences (PDS) which was not present in the

FDS situation is that the life path of the simulated biography may change depending on the outcome of previous decisions. On the other hand, in a state space where decision steps are arranged in a fixed sequence, all biographies experience the same life paths (decision processes) although outcomes of previous events may quantitatively modulate later decisions.

3.2.3 Mixed Case: FDS and PDS

The decision dependencies discussed in Sections 3.2.1 and 3.2.2 above describe the basic types of decision processes which may arise in the state space of a longitudinal microsimulation model. Most MSM models will consist of mixed decision structures with the fixed FDS segment containing M_t ($< S_t$) decision steps preceding the random PDS segment.

In order to represent the set of all possible decision paths $\underset{\sim}{U}_t$ in the mixed case, we first define its two components: $\underset{\sim}{U}_t^F = \left\{ \underset{\sim}{D}_{t1}^F \right\} = \left\{ (D_{t1}, \ldots, D_{tM_t}) \right\}$ denoting the possible paths (only one) in the FDS segment and $\underset{\sim}{U}_t^P = \left\{ \underset{\sim}{D}_{t1}^P, \ldots, \underset{\sim}{D}_{tQ_t}^P \right\}$ denoting the possible paths on the PDS segment. Then, our set of possible decision paths at time t may be written as the cartesian product of $\underset{\sim}{U}_t^F$ and $\underset{\sim}{U}_t^P$: $\underset{\sim}{U}_t = (\underset{\sim}{U}_t^F \times \underset{\sim}{U}_t^P)$.

3.3 Simulated Decision Outcomes

In this section we examine in greater detail how the components of the endogenous outcome vector $\underset{\sim}{\beta}_{it}$ are created for a simulated biography i at time t. Recall that $\underset{\sim}{\beta}_{it}$ is one particular realization of a decision path $\underset{\sim}{D}_{tq} \in \underset{\sim}{U}_t$ using the random number stream $\underset{\sim}{r}_{it} = (r_{it1}, \ldots, r_{itS_t})$ drawn at time t to advance the biography from the achieved state $\underset{\sim}{\beta}_{i,t-1}$ to the new state $\underset{\sim}{\beta}_{it}$.

Each decision function (step) D_{tk} in the decision set $U_t = \{ D_{t1}, \ldots, D_{tS_t} \}$ is a random variable. The decision function D_{tk} for decision k at time t takes as argument a random uniform variate $r_{itk} \in U[0, 1]$, and given the outcome of earlier decision steps $\underset{\sim}{\beta}_{it[1:k-1]} = (\beta_{it}^{(1)}, \ldots, \beta_{it}^{(k-1)})$, maps r_{itk} onto a point in the decision space X_{itk}. Probabilities of transition from a state in $X_{t,k-1}$ to states in the decision space X_{itk} are obtained from the transition density function $f_{D_{tk}/D_{t,k-1}}(\cdot)$. All probability functions $f_{D_{tk}/D_{t,k-1}}(\cdot)$, $t = 1, \ldots, n$, $k = 1, \ldots, S_t$, used enter the decision process as components of the exogenous information sets $\underset{\sim}{\alpha}_t$. In addition, transition densities $f_{D_{tk}/D_{t,k-1}}(\cdot)$ at time t may be altered by endogenous outcomes $\underset{\sim}{\beta}_{is}$, $s = 1, \ldots, t-1$, obtained while simulating earlier time points $s = 1, \ldots, t-1$.

The dependence of the conditional transition pdf and cdf on both the exogenous information set $\underset{\sim}{\alpha}_t$ and the realized endogenous states $\underset{\sim}{\beta}_{i,t-1}$ and $\underset{\sim}{\beta}_{it[1:k-1]}$ is captured in the definition below.

Definition 3.3: Conditional pdf and cdf of transition from the source decision space $X_{t(k-1)}$ to the destination decision space X_{tk}

$$f_{D_{tk}/D_{t,k-1}}(\cdot) = f_{D_{tk}/D_{t,k-1}}(\cdot\,; \underset{\sim}{\theta}_{itk}(\alpha_t, \underset{\sim}{\beta}_{i,t-1}, \underset{\sim}{\beta}_{it[1:k-1]}))$$

$$F_{D_{tk}/D_{t,k-1}}(\cdot) = F_{D_{tk}/D_{t,k-1}}(\cdot\,; \underset{\sim}{\theta}_{itk}(\underset{\sim}{\alpha}_t, \underset{\sim}{\beta}_{i,t-1}, \underset{\sim}{\beta}_{it[1:k-1]}))$$

where $\underset{\sim}{\theta}_{itk}(\alpha_t, \underset{\sim}{\beta}_{i(t-1)}, \underset{\sim}{\beta}_{it[1:k-1]})$ are parameters which impact the shape and range of the transition pdf and cdf faced by biography i in executing decision steps D_{tk}.

We are now poised to formulate the concept of a simulated outcome for biography i at time t using the defintion of the decision function. The labelling of the random numbers used to make each decision is crucial in connecting the outcome of a decision function with the simulated outcome of a biography. In Definition 3.4 below the original Definition 3.2 of the decision function D_{tk} introduced in Section 3.1 is modified by replacing the Monte-Carlo draw u with r_{itk} to explictly link the decision outcome $\beta_{it}^{(k)}$ for biography i at time t with the decision function D_{tk} and its input r_{itk}.

Definition 3.4: Decision Step D_{tk} and Decision Outcome $\beta_{it}^{(k)}$
Given a source decision state $x_{t,k-1} \in X_{t,k-1}$ and the cdf parameter $F_{D_{tk}/D_{t(k-1)}}(\cdot)$ to the decision function D_{tk}, the decision outcome $\beta_{it}^{(k)}$ in the destination space X_{tk} obtained using the Monte-Carlo draw r_{itk} is described by the following probability map:

$$D_{tk}(r_{itk}; F_{D_{tk}/D_{t,k-1}}): \quad r_{itk} \in U[0,1] \quad \rightarrow \quad \beta_{it}^{(k)} \in X_{tk}$$

$$\text{such that} \quad \beta_{it}^{(k)} = F^{-1}_{D_{tk}/D_{t,k-1}}(r_{itk}), \quad \text{if } X_{tk} \text{ continous}$$

$$\beta_{it}^{(k)} = \inf\{x \in X_{tk} \mid F_{D_{tk}/D_{t,k-1}}(x) \geq r_{itk}\}, \quad \text{if } X_{tk} \text{ discrete}$$

Moreover, the relationship between the labelled random variate r_{itk} and the decision outcome $\beta_{it}^{(k)}$ for biography i at time t may be further stressed by expressing the definition above more compactly using the functional notation given in Definition 3.5 below.

Definition 3.5: Outcome $\beta_{it}^{(k)}$ of Decision Step D_{tk} for Biography i at time t

$$\beta_{it}^{(k)} = D_{tk}\left(r_{itk}; F_{D_{tk}/D_{t,k-1}}(\cdot\,; \underset{\sim}{\theta}_{itk}(\underset{\sim}{\alpha}_t, \underset{\sim}{\beta}_{i,t-1}, \underset{\sim}{\beta}_{it[1:k-1]}))\right). \qquad (3.1)$$

Having developed the concept of an individual outcome $\beta_{it}^{(k)}$ for a particular decision step D_{tk}, we now formulate an expression for the complete endogenous outcome state vector $\underset{\sim}{\beta}_{it}$ for individual i at time t. Simulation of biography i at time t requires a number of decision steps to be taken in a

sequence. The ordering of this sequence may be fixed (FDS), yielding $\underset{\sim}{U}_t = \{\underset{\sim}{D}_{t1}\}$, or the ordering may be subject to permutation (PDS) in which case the set of all possible decision sequence configurations at time t contains more than one element $\underset{\sim}{U}_t = \{\underset{\sim}{D}_{t1}, \ldots, \underset{\sim}{D}_{tQ_t}\}, Q_t > 1$.

When the ordering of decision steps is fixed, the number of random variates required to simulate any biography at time t is also fixed. In cases where the dependency of decision steps follows the FDS type of structure, each ordered element of $\underset{\sim}{r}_{it} = (r_{it1}, \ldots, r_{itS_t})$ corresponds exactly in position with each element of the decision sequence $\underset{\sim}{D}_{t1} = (\mathrm{D}_{t1}, \ldots, \mathrm{D}_{tS_t})$. A functional representation for the endogenous outcome state vector $\underset{\sim}{\beta}_{it}$ for biography i at time t, based on Definition 3.5, is given below which directly links the decision function sequence $\underset{\sim}{D}_{t1} = (\mathrm{D}_{t1}, \ldots, \mathrm{D}_{tS_t})$ with its random input stream $\underset{\sim}{r}_{it} = (r_{it1}, \ldots, r_{itS_t})$ and cdf parameter vector $\underset{\sim}{F}_{t[1:S_t]} = (F_{D_{t1}}, \ldots, F_{\mathrm{D}_{tk}/\mathrm{D}_{t,k-1}}, \ldots, F_{\mathrm{D}_{tS_t}/\mathrm{D}_{t,S_t-1}})$. For ease of notation, the cdf parameter will be written simply as $\underset{\sim}{F}_{t[1:S_t]} = (F_{D_{t1}}, \ldots, F_{\mathrm{D}_{tk}}, \ldots, F_{\mathrm{D}_{tS_t}})$.

Definition 3.6 Outcome Vector $\underset{\sim}{\beta}_{it}$ in FDS Case

When the ordering of the decision steps is non-permutable, there exists a one-to-one correspondence between the elements of the random variate stream $\underset{\sim}{r}_{it} = (r_{it1}, \ldots, r_{itS_t})$ drawn to simulate biograhy i at time t and the decision steps in $\underset{\sim}{D}_{t1} = (\mathrm{D}_{t1}, \ldots, \mathrm{D}_{tS_t})$. The complete endogenous outcome state vector $\underset{\sim}{\beta}_{it}$ simulated at time t for biograhy i may be expressed functionally (by Definition 3.5) as follows:

$$\underset{\sim}{\beta}_{it} = (\mathrm{D}_{t1}(r_{it1}; \mathrm{F}_{\mathrm{D}_{t1}}(\cdot)), \ldots, \mathrm{D}_{tS_t}(r_{itS_t}; \mathrm{F}_{\mathrm{D}_{tS_t}}(\cdot)))$$

$$= \underset{\sim}{D}_{t[1:S_t]}(\underset{\sim}{r}_{it[1:S_t]}, \underset{\sim}{\mathrm{F}}_{\mathrm{D}_t[1:S_t]}(\cdot))$$

In the case where the ordering of decision steps is open to permutation (PDS), this one-to-one correspondence between $\underset{\sim}{r}_{it}$ and decision sequences in $\underset{\sim}{U}_t = \{\underset{\sim}{D}_{t1}, \ldots, \underset{\sim}{D}_{tQ_t}\}$ will no longer hold throughout the whole sequence because decision steps in the sequence may vary for different biographies. In this situation, different biographies may randomly follow different decision paths.

In most MSM situations, however, both FDS and PDS type segments will hold in the decision space. Here, permutable decision sequences (PDS segment) arise after the fixed decision path (FDS segment) giving rise to the mixed decision path structures discussed in Section 3.2. For instance, in the example provided in Section 3.2.2, the set of all possible decision sequences was $\underset{\sim}{U}_t = \{\underset{\sim}{D}_{t1}, \underset{\sim}{D}_{t2}, \underset{\sim}{D}_{t3}\} = \{(\mathrm{D}_{t1}, \mathrm{D}_{t2}), (\mathrm{D}_{t1}, \mathrm{D}_{t2}, \mathrm{D}_{t3}), (\mathrm{D}_{t1}, \mathrm{D}_{t2}, \mathrm{D}_{t4})\}$.

In all three possible decision path occurrences, decisions D_{t1} and D_{t2} occur commonly in the first segment of all decision paths. The size of the random variate stream $\underset{\sim}{r}_{it} = (r_{it1}, r_{it2}, \ldots)$ used to advance biography i in the interval $(t-1, t]$ will vary depending on which decision path $\underset{\sim}{D}_{tq}$, $q = \{1, 2, 3\}$, is

followed by the simulated biography. However, among all three realizable decision paths, the first two random variates drawn, r_{it1} and r_{it2}, will always correspond to the first two decision steps D_{t1} and D_{t2}.

More generally, the decision paths which may arise in a mixed state space at time t were represented by $\underset{\sim}{U}_t = (\underset{\sim}{U}_t^F \times \underset{\sim}{U}_t^P)$ (see Section 3.2.3). For all biographies simulated, the first $M_t < S_t$ decision steps will remain constant. This enables us to represent the cdf of conditional supports for the decision spaces $X_{tk}, k=1,\ldots,m_t$, by $\underset{\sim}{F}_t^F = \underset{\sim}{F}_{t[1:M_t]} = (F_{D_{t1}}, \ldots, F_{D_{tk}}, \ldots, F_{D_{tM_t}})$. Similarly, we allow $\underset{\sim}{r}_{it}^F = \underset{\sim}{r}_{it[1:M_t]} = (r_{it1}, \ldots, r_{itM_t})$ to be that sub-stream of $\underset{\sim}{r}_{it}$ used to determine the outcome of the decision steps in $\underset{\sim}{U}_t^F$. We represent the decision outcomes for biography i at time t obtained from the FDS segment of the decision path by $\underset{\sim}{\beta}_{it}^F = \underset{\sim}{\beta}_{it[1:M_t]} = (\beta_{it}^{(1)}, \ldots, \beta_{it}^{(M_t)})$. Definition 3.6 allows these outcomes to be written in terms of the Monte-Carlo stream $\underset{\sim}{r}_{it}^F$ and cdf parameter $\underset{\sim}{F}_t^F$ as follows:

$$\begin{aligned}
\underset{\sim}{\beta}_{it}^F &= (D_{t1}(r_{it1}, F_{D_{t1}}(\cdot)), \ldots, D_{tM_t}(r_{itM_t}, F_{D_{t1}}(\cdot))) \\
&= \underset{\sim}{D}_{t[t:M_t]}(\underset{\sim}{r}_{it[1:M_t]}, \underset{\sim}{F}_{D_{t[1:M_t]}}(\cdot)) \\
&= \underset{\sim}{D}_t^F(\underset{\sim}{r}_{it}^F, \underset{\sim}{F}_t^F)
\end{aligned}$$

Meanwhile, outcomes in the complement of $\underset{\sim}{\beta}_{it}^F$ denoted by $\underset{\sim}{\beta}_{it}^P$ $\left(\underset{\sim}{\beta}_{it} = (\underset{\sim}{\beta}_{it}^F, \underset{\sim}{\beta}_{it}^P)\right)$ arise from decisions from the permutable PDS segment of the decision path taken by the biography. Decisions in the permutable segment of the decision sequence realized for biography i depend on outcomes of decision steps taken by the biography in earlier decisions and becomes completely specified only once the last decision in the path is taken.

This implies that the length of $\underset{\sim}{r}_{it}^P$ is random and depends on the decision path realized in $\underset{\sim}{U}_t^P$. Therefore, the length of the entire Monte-Carlo stream $\underset{\sim}{r}_{it} = (\underset{\sim}{r}_{it}^F, \underset{\sim}{r}_{it}^P)$ needed to simulate a biography at time t is also random. The complete outcome vector may be expressed in terms of decision sequences and their arguments as

$$\begin{aligned}
\underset{\sim}{\beta}_{it} &= (\underset{\sim}{\beta}_{it}^F, \underset{\sim}{\beta}_{it}^P) \\
&= \left(\underset{\sim}{D}_{t[1:M_t]}(\underset{\sim}{r}_{t[1:M_t]}), D_t^P(\underset{\sim}{r}_{it}^P)\right).
\end{aligned}$$

4. POLICY COMPARISONS AND ACROSS-POLICY MONTE-CARLO EFFECTS

An important use of longitudinal MSM is in the area of policy development and strategic planning. MSM affords the researcher the ability to model the essential determinants of random real world

phenomena under various assumptions. For example, in the context of POHEM microsimulation, the researcher is interested in the health outcome of a hypothetic Canadian population (constructed from observations of the actual population) under various health policy scenarios (Wolfson and Berthelot, 1992). The researcher wishes to study how health measures and health costs are affected by public policy initiatives aimed at discouraging smoking or lowering the cholesterol level for instance.

4.1 Concept and Representation of a Policy in MSM

The micro-level theory for longitudinal microsimulation developed in Section 3 allows us to describe the simulation process for individual states $\beta_{it}^{(k)}$ in $\underset{\sim}{\beta}_{it} = (\beta_{it}^{(1)}, \ldots, \beta_{it}^{(S_t)})$ at an elemental level as follows:

$$\beta_{it}^{(k)} = \mathrm{D}_{tk}\left(r_{itk};\ F_{D_{tk}|D_{t,k-1}=\beta_{it}^{(k-1)}}(\cdot;\ \underset{\sim}{\theta}_{itk}(\underset{\sim}{\alpha}_t, \underset{\sim}{\beta}_{i,t-1}, \underset{\sim}{\beta}_{it[1:k-1]}))\right) \tag{4.1}$$

for $k=1, \ldots, S_t$.

As discussed earlier, the decision function D_{tk} maps the uniform random variate r_{itk} onto a point $\beta_{it}^{(k)}$ in the decision space X_{tk}. This mapping is performed with the help of the transition cdf $F_{D_{tk}|D_{t,k-1}}(\cdot;\ \theta_{itk}(\underset{\sim}{\alpha}_t, \underset{\sim}{\beta}_{i,t-1}, \underset{\sim}{\beta}_{it[1:k-1]}))$ parameterized by the exogenous information set $\underset{\sim}{\alpha}_t$ and the outcome states $\underset{\sim}{\beta}_{i,t-1}$ and $\underset{\sim}{\beta}_{it[1:k-1]}$ achieved earlier.

A policy in the mathematical model for MSM proposed in equation (4.1) is represented by variables in the exogenous state vector $\underset{\sim}{\alpha}_t$. This vector describes all external assumptions to the MSM model such as the effect of risk factors (egs. smoking, cholesterol), socio-economic factors, genetic disposition and the like on the survival distribution of age-sex based biographies which eventually quantify the structural connections over the state space. Clearly, policy interventions directed at changing smoking and cholesterol levels through behavioral change (eg. exercise, diet) represent changes to the exogenous state vector $\underset{\sim}{\alpha}_t$, $t \geq 1$.

Changes in outcome states $\underset{\sim}{\beta}_{it}$ are the result of quantitative changes induced on the 'laws of motion' guiding the MSM and do not cause structural changes in the model. By this we mean that the state space in not structurally changed by policy alterations in the sense that new states are not added or old states deleted at any time in the state space. Nor are flow connections across time between states altered by policy changes; only the probabilities of transitions among states are eventually effected by policy changes. It is important to understand that structural changes to the state space in the form of addition/deletion of states and flow connections represent changes to the 'laws of motion' governing the system. Such changes alter the structure of the model and should not be confused with policy changes. Therefore, policy changes hold the structure of the model constant but alter the likelihoods of transitions

to connected states.

4.2 Types of Policy Change

When comparing two policies for each simulated biography, represented by $\underset{\sim}{\alpha} = (\underset{\sim}{\alpha}_1, \ldots, \underset{\sim}{\alpha}_t, \ldots, \underset{\sim}{\alpha}_\infty)$ and $\alpha' = (\underset{\sim}{\alpha}'_1, \ldots, \underset{\sim}{\alpha}'_t, \ldots, \underset{\sim}{\alpha}'_\infty)$, the researcher is interested in studying the impact differences in the two policy assumptions have on the corresponding simulated outcomes $\underset{\sim}{\beta}_i = (\underset{\sim}{\beta}_{i1}, \ldots, \underset{\sim}{\beta}_{it}, \ldots, \underset{\sim}{\beta}_{iT_i})$ and $\beta'_i = (\underset{\sim}{\beta}'_{i1}, \ldots, \underset{\sim}{\beta}'_{it}, \ldots, \underset{\sim}{\beta}'_{iT'_i})$. The representation for $\underset{\sim}{\alpha}'$ above assumes that policy interventions are made at all times in the simulation horizon.

Other types of policy change may be of possible interest to the researcher. For example, the case of policy intervention only at time t_0 ($\alpha' = (\underset{\sim}{\alpha}_1, \ldots, \underset{\sim}{\alpha}_{t_0-1}, \underset{\sim}{\alpha}'_{t_0}, \alpha_{t_0+1} \ldots, \underset{\sim}{\alpha}_\infty)$) is a special case of the policy change mentioned above. Here, change to the exogenous conditions facing the biography occurs only at time t_0; conditions at all other times before or after remain the same as before. Yet, another variety of policy intervention is when interventions affect exogenous conditions at time t_0 and also thereafter for all times but not before t_0; i.e. $\alpha'_i = (\underset{\sim}{\alpha}_1, \ldots \underset{\sim}{\alpha}_{t_0-1}, \underset{\sim}{\alpha}'_{t_0}, \underset{\sim}{\alpha}'_{t_0+1}, \ldots, \underset{\sim}{\alpha}'_\infty)$. This situation also leads to a similar pattern of outcomes as in the case of intervention at time t_0 only.

4.3 Confounding of Policy & Monte-Carlo Effects

In order to investigate the impacts of policy change on the simulated outcome $\beta_{it}^{(k)}$, consider the consequences of a policy shift from $\underset{\sim}{\alpha}$ to $\underset{\sim}{\alpha}'$. To maintain simplicity in the exposition, without sacrificing generality, we assume that the policy change $\underset{\sim}{\alpha}$ involves only a change in policy at time t so that if $\underset{\sim}{\alpha} = (\underset{\sim}{\alpha}_1, \ldots, \underset{\sim}{\alpha}_t, \ldots, \underset{\sim}{\alpha}_\infty)$ represents the baseline policy, then after policy shift $\alpha' = (\underset{\sim}{\alpha}_1, \ldots, \underset{\sim}{\alpha}'_t, \ldots, \underset{\sim}{\alpha}_\infty)$ represents the modified policy.

Using the results established in Section 3.3, the different outcomes for decision k at time t under the two policies may be represented by the following models:

Policy I ($\underset{\sim}{\alpha}_i$):

$$\beta_{it}^{(k)} = \mathrm{D}_{tk}\left(r_{itk};\ F_{D_{tk}|D_{t,k-1}=\beta_{it}^{(k-1)}}(\cdot;\ \underset{\sim}{\theta}_{itk}(\underset{\sim}{\alpha}_t, \underset{\sim}{\beta}_{i,t-1}, \underset{\sim}{\beta}_{it[1:k-1]}))\right), \quad \begin{matrix} i = 1, \ldots, n; \\ t = 1, \ldots, T_i; \\ k = 1, \ldots, S_t. \end{matrix} \tag{4.2}$$

Policy II ($\underset{\sim}{\alpha}'_i$):

If the decision structure is fixed (FDS):

$$\beta_{it}^{\prime(k)} = \mathrm{D}_{tk}\left(s_{itk};\ F_{D_{tk}|D_{t,k-1}=\beta_{it}^{\prime(k-1)}}(\cdot;\ \underset{\sim}{\theta}'_{itk}(\underset{\sim}{\alpha}'_t, \underset{\sim}{\beta}_{i,t-1}, \beta'_{it[1:k-1]}))\right), \tag{4.3}$$

If the decision structure is permutable (PDS):

$$\beta_{it}^{(k')} = \mathrm{D}_{tk'}\left(s_{itk};\ F_{D_{tk}|D_{t,k'-1}=\beta_{it}^{(k'-1)}}(\cdot;\ \underset{\sim}{\theta}_{itk'}(\underset{\sim}{\alpha}_t', \underset{\sim}{\beta}_{i,t-1}, \beta_{it[1:k'-1]}'))\right) \tag{4.4}$$

It is important to note that both r_{itk} and s_{itk} are uniform [0, 1] Monte-Carlo draws drawn to simulate policy outcomes for biographies $i = 1, \ldots, n$ under policy (1) and policy (2). Since policy runs are performed separately, r_{itk} and s_{itk} are drawn independently under both policy runs and furthermore r_{itk} will be usually different from s_{itk}. Here, i simply labels the order in which biographies are generated under the two policies.

The effects of a shift in policy from $\underset{\sim}{\alpha}_{it}$ to $\underset{\sim}{\alpha}_{it}'$ on the outcome $\underset{\sim}{\beta}_{it}$ are altogether different depending on whether the state space gives rise to a fixed (FDS) or permutable (PDS) decision path structure.

FDS Case

Recall that in the FDS situation the set of all possible decision sequence configurations $\underset{\sim}{U}_t$ at each time t has only one element: $\underset{\sim}{D}_t = \{D_{t1}, \ldots, D_{tQ_t}\}$, $Q_t = 1$. This implies that the length of the random stream $\underset{\sim}{r}_{it}$, used to simulate biogaphy i at moment t is fixed and each element of $\underset{\sim}{r}_{it} = (r_{it1}, \ldots, r_{itS_t})$ has a one-to-one exact correspondence with fixed ordered decision steps in $\underset{\sim}{D}_{t1} = (\mathrm{D}_{t1}, \ldots, \mathrm{D}_{tS_t})$ and the kth random variate in $\underset{\sim}{r}_{it}$, namely r_{itk}, serves as input to the kth decision D_{tk}: $\beta_{it}^{(k)} = \mathrm{D}_{tk}(r_{itk})$.

Although the order and length of the decision path remains fixed in the FDS case, decision steps occuring later in the path, say $\mathrm{D}_{tk}, k > 1$, depend conditionally on outcomes of previous decisions $\underset{\sim}{\beta}_{it[1:k-1]} = \underset{\sim}{D}_{t[1:k-1]}(\underset{\sim}{r}_{it[1:k-1]})$ through their effect of their outcomes on the parameters of the transition cdf $F_{D_{tk}|D_{t,k-1}}(\cdot;\ \underset{\sim}{\theta}_{itk}(\underset{\sim}{\alpha}_t, \underset{\sim}{\beta}_{i,t-1}, \underset{\sim}{\beta}_{it[1:k-1]}))$ supporting the decision space X_{tk} of D_{tk}.

Comparing expressions (4.2) and (4.3) representing the outcome from the same decision step D_{tk} under the two policies $\underset{\sim}{\alpha}_t$ and $\underset{\sim}{\alpha}_t'$, one observes that changes in the outcomes $\beta_{it}^{(k)}$ and $\beta_{it}^{\prime(k)}$ are the result of two influences:

i) Policy effect represented by change in policy from $\underset{\sim}{\alpha}_t$ to $\underset{\sim}{\alpha}_t'$ and

ii) Across-policy Monte-Carlo effect due to using different random variate draws $r_{itk} \neq s_{itk}$.

From these observations, it is clear that the difference in outcomes $\beta_{it}^{\prime(k)} - \beta_{it}^{(k)}$ under the two policy scenarios is confounded by both a policy effect and an accompanying across-policy Monte-Carlo effect. Ideally, one would like to observe the impact of policy change on microsimulation output independently of such Monte-Carlo contamination. In Section 4.5, an optimal biography generation design is proposed which blocks out such Monte-Carlo contamination (confounding).

PDS and Mixed Case

In the PDS situation, more than one ordering among decision steps can arise at each time t: $\underset{\sim}{U}_t = \{\underset{\sim}{D}_{t1}, \ldots, \underset{\sim}{D}_{tQ_t}\}$, $Q_t > 1$. The effect of policy change α_t' on the outcome $\beta_{it}^{(k)}$ is more complex. Here, results from earlier outcomes $\beta_{it[1:k'-1]}'$ do not just quantitatively modulate later decisions but may alter the subsequent decision path taken by the biography. Therefore, the decision steps taken at time t and later under the two policies may be different. The propensity of the decision step to change is reflected by the prime in $D_{tk'}$. Consequently, the parity between the outcome $\beta_{it}^{(k)} = D_{tk}(r_{itk})$ under policy $\underset{\sim}{\alpha}_t$ and the outcome $\beta_{it}^{(k')} = D_{tk'}(r_{itk})$ under policy $\underset{\sim}{\alpha}_t'$ with respect to the Monte-Carlo draw r_{itk} is broken.

Decision outcomes under the two policy scenarios are no longer comparable because the decision spaces X_{tk} and $X_{tk'}$ are no longer the same. The random variate stream $\underset{\sim}{r}_{it}$ now varies depending on the decision path $\underset{\sim}{D}_{tq} \in \underset{\sim}{U}_t$ selected by the biography and the one-to-one exact correspondence between the random variates and decision steps no longer persists across policies.

In mixed decision sequences, a fixed (FDS) sub-sequence is followed by a permutable (PDS) one (see Section 3.3) and the one-to-one correspondence between the positioning of Monte-Carlo draws r_{itk} in the random number stream $\underset{\sim}{r}_{it}$ and decision steps D_{tk} holds only in the FDS segment of $\underset{\sim}{U}_t = (\underset{\sim}{U}_t^F \times \underset{\sim}{U}_t^P)$. Recall that outcomes arising from mixed decision sequences were represented as

$$\underset{\sim}{\beta}_{it} = (\underset{\sim}{\beta}_{it}^F, \underset{\sim}{\beta}_{it}^P)$$

$$= \left(\underset{\sim}{D}_{t[1:M_t]}(\underset{\sim}{r}_{t[1:M_t]}), D_t^P(\underset{\sim}{r}_{it}^P)\right).$$

where $\underset{\sim}{\beta}_{it}^F = \underset{\sim}{\beta}_{it[1:M_t]} = (\mathrm{D}_{t1}(r_{it1}), \ldots, \mathrm{D}_{tM_t}(r_{itM_t})) = \underset{\sim}{D}_{t[1:M_t]}(\underset{\sim}{r}_{it[1:M_t]}) = \underset{\sim}{D}_t^F(\underset{\sim}{r}_{it}^F)$ represents outcomes for the FDS segment of the decision sequence $\underset{\sim}{D}_{t[1:M_t]}$ and $\underset{\sim}{\beta}_{it}^F$ represents outcomes obtained upon executing decision steps in the random PDS component $\underset{\sim}{D}_t^P$. If the Monte-Carlo stream $\underset{\sim}{r}_{it} = (\underset{\sim}{r}_{it}^F, \underset{\sim}{r}_{it}^P)$ is used to simulate $\underset{\sim}{\beta}_{it}$, then the one-to-one correspondence between the decision functions D_{tk} and their inputs r_{itk} holds only among elements of $\underset{\sim}{D}_t^F = \underset{\sim}{D}_{t[1:M_t]} = (\mathrm{D}_{t1}, \ldots, \mathrm{D}_{tM_t})$ and $\underset{\sim}{r}_{it}^F = \underset{\sim}{r}_{it[1:M_t]} = (r_{it1}, \ldots, r_{itM_t})$.

Therefore, the biography generation design directed at blocking out Monte-Carlo effects in policy comparisons may be applied to the fixed decision path segment $\underset{\sim}{D}_{t[1:M_t]}$ of all decision paths in $\underset{\sim}{U}_t$.

4.4 Optimal Biography Generation Design

From Section 4.4 it can be seen that unless control is exercised on biography generation policy effects become confounded with Monte-Carlo effects. Due to this confounding nature of policy effects and Monte-Carlo influences, it becomes difficult to evaluate the direct impact of policy change in

simulation outcomes - whether differences in response are due to policy changes ($\underset{\sim}{\alpha}$ vs. $\underset{\sim}{\alpha}'$) or are they the result of uncontrolled differences in the random numbers (r_{itk} vs s_{itk}) used. From a design perspective this is undesirable because inadvertent changes in the random numbers used in performing different policy scenarios blur the real effects of policy change.

Optimally, if control over biography generation was possible, one would wish to produce microsimulation output for the same set of biographies under the two policies $\underset{\sim}{\alpha}$ and $\underset{\sim}{\alpha}'$ which was free from confounding Monte-Carlo effects. Expressions (4.2) to (4.4) suggest that a design (call it D2 design) which would block out Monte-Carlo influences across policy scenarios is one which used the same random numbers r_{itk}, $i = 1, \ldots, n$; $t = 1, \ldots, T_i$; $k = 1, \ldots, S_t$, to simulate the common decisions under both policy scenarios $\underset{\sim}{\alpha}$ and $\underset{\sim}{\alpha}'$. This would ensure that $r_{itk} = s_{itk}, \forall_{i,t,k}$. Under the D2 design, the simulated outcome for decision k at time t would be represented by the following models:

Policy I (α_t):

$$\beta_{it}^{(k)} = \mathrm{D}_{\mathrm{tk}}\left(r_{itk};\ F_{D_{tk}|D_{t,k-1}=\beta_{it}^{(k-1)}}(\cdot;\ \underset{\sim}{\theta}_{itk}(\underset{\sim}{\alpha}_t, \underset{\sim}{\beta}_{i,t-1}, \underset{\sim}{\beta}_{it[1:k-1]}))\right), \quad \begin{matrix} i = 1, \ldots, n; \\ t = 1, \ldots, T_i; \\ k = 1, \ldots, S_t. \end{matrix} \tag{4.5}$$

Policy II (for FDS segment) (α_t'):

$$\beta_{it}^{(k)\prime} = \mathrm{D}_{\mathrm{tk}}\left(r_{itk};\ F_{D_{tk}|D_{t,k-1}=\beta_{it}^{(k-1)}}(\cdot;\ \underset{\sim}{\theta}'_{itk}(\underset{\sim}{\alpha}'_t, \underset{\sim}{\beta}_{i,t-1}, \underset{\sim}{\beta}'_{it[1:k-1]}))\right), \quad \begin{matrix} i = 1, \ldots, n; \\ t = 1, \ldots, T_i; \\ k = 1, \ldots, S_t. \end{matrix} \tag{4.6}$$

The use of the same Monte-Carlo draw r_{itk} in both policy scenarios for the FDS segment removes influences on the generated outcome states due to differences in the random numbers used. This ensures that in a Monte-Carlo sense the same "genetic" biography is observed under different policies and allows the effect of policy change to be measured in the absence of confounding Monte-Carlo effects.

It is important to note that based on the discussion in Section 4.4, blocking out Monte-Carlo effects across policies is meaningful only among outcomes generated from decision steps contained in FDS types of decision sequences or in the case of mixed sequences from decision steps located in the fixed segment of the sequence. However, the benefits of reducing across-policy Monte-Carlo variation in the fixed segment of the decision path will also carry over to the permutable component of the decision path because outcomes $\underset{\sim}{\beta}_{i,t-1}$ and $\underset{\sim}{\beta}_{it[1:k-1]}$ obtained from the fixed decision sub-sequence parameterize the transition cdf $F_{D_{tk}|D_{t,k-1}}(\cdot;\ \underset{\sim}{\theta}_{itk}(\underset{\sim}{\alpha}_t, \underset{\sim}{\beta}_{i,t-1}, \underset{\sim}{\beta}_{it[1:k-1]}))$ of decision steps in the PDS component of the decision path.

Controlling Monte-Carlo fluctuations across policies in the FDS segment of the decision path will

contribute a lower level of across-policy Monte-Carlo contamination to the stochastic profile of transition cdfs in the permutable section of the decision path.

Some schemes for implementing the proposed D2 design are given in Appendix 1.

5. STATISTICAL MODEL FOR MSM AND ITS USE IN POLICY COMPARISONS

In this section, based on the theory and concepts developed in the previous two sections, we posit a statistical model representation for generated MSM outcomes and develop a methodology for estimating and testing the impact of policy from data generated under different policy scenarios. Once an appropriate statistical representation for MSM output has been specified, the gains from the proposed D2 biography generation design of Section 3.3 may be clearly demonstrated.

5.1 Statistical Model for MSM Output

MSM output in its most disaggregate form is available for each biography i over its simulated life as a vector of endogenous outcomes $\underset{\sim}{\beta}_i = (\underset{\sim}{\beta}_{i1}, \ldots, \underset{\sim}{\beta}_{iT_i})$. The simulated outcome vector for the biography at time t, $\underset{\sim}{\beta}_{it} = (\beta_{it}^{(1)}, \ldots, \beta_{it}^{(S_t)})$, contains generated values for many characteristics whose k th element $\beta_{it}^{(k)}$ denotes the observed value for characteristic k (e.g. medical cost). The time-aggregated outcome for characteristic k, $y_{i.k}$, may be functionally represented in terms of biography i's time-wise simulated values $\underset{\sim}{\beta}_i = (\underset{\sim}{\beta}_{i1}, \ldots, \underset{\sim}{\beta}_{iT_i})$. In the simplest case, each time aggregated element $y_{i.k}$ will be a simple aggregation of time-wise outcomes $\beta_{it}^{(k)}, t = 1, \ldots, T_i$, for characteristic k. In this situation, keeping in mind that $y_{itk} \equiv \beta_{it}^{(k)}$ represents for instance the medical cost incurred by individual i at time t; we may write

$$y_{i.k} = \sum_{t=1}^{T_i} \beta_{it}^{(k)} \tag{5.1}$$

Making use of the functional representation of $\beta_{it}^{(k)}$ in terms of input r_{itk} and cdf parameter $F_{D_{tk}/D_{t(k-1)}}(\cdot)$ given by Definition 3.4, the time-aggregate total medical cost (characteristic k) of individual i may be expressed as follows:

$$y_{i.k} = \sum_{t=1}^{T_i} y_{itk} = \sum_{t=1}^{T_i} D_{tk}\left(r_{itk}; F_{D_{tk}|D_{t,k-1}}(\cdot\,; \underset{\sim}{\theta}_{itk}(\underset{\sim}{\alpha}_t, \underset{\sim}{\beta}_{i,t-1}, \underset{\sim}{\beta}_{it[1:k-1]}))\right) \tag{5.2}$$

Relationships among the arguments to $D_{tk}(r_{itk}; F_{D_{tk}|D_{t,k-1}}(\cdot\,; \underset{\sim}{\theta}_{itk}(\underset{\sim}{\alpha}_t, \underset{\sim}{\beta}_{i,t-1}, \underset{\sim}{\beta}_{it[1:k-1]})))$ are very complex and the whole microsimulation model is expressible only as a computer program; an analytical

representation is too complicated to derive. In devising a comparison methodology we need to first abstract the essential features of microsimulation output (using the results of Section 3) and model them in the simplest way convenient to estimate and test statistically.

Allow superscript p to denote the policy in effect, t to represent the simulation period $(t-1, t]$, k to denote the characteristic simulated from decision step D_{tk}, and i to represent the biography simulated using the Monte-Carlo draw r_{itk}. We posit the following additive statistical model for the outcome $y_{itk}^{(p)}(r_{itk})$:

$$y_{itk}^{(p)}(r_{itk}) = \mu_{tk}^{(p)} + \gamma_{itk} + \epsilon_{itk}^{(p)} \tag{5.4}$$

where the model components having the following interpretations:

$\mu_{tk}^{(p)} \equiv \mu_{tk}^{(p)}(F_{D_{tk}}(\cdot\,; \underset{\sim}{\theta}(\underset{\sim}{\alpha}_t^{(p)})))$:

Mean level of microsimulation output $y_{itk}^{(p)}$ for characteristic k under policy p due to the effect of policy $\underset{\sim}{\alpha}_t^{(p)}$ on $F_{D_{tk}}(\cdot)$.

$\gamma_{itk} \equiv \gamma_{itk}(r_{itk}) \sim (0, \sigma^2_{\gamma_{tk}})$: Random Monte-Carlo effect induced by the selection of the random Monte-Carlo draws r_{itk} ($r_{itk} \sim$ U [0, 1]) to execute decision D_{tk} for biography i.

$\epsilon_{itk}^{(p)} \equiv \epsilon_{itk}^{(p)}\left(F_{D_{tk}}(\cdot\,; \underset{\sim}{\theta}_{itk}(\underset{\sim}{\beta}_{i,t-1}, \underset{\sim}{\beta}_{it[1:k-1]})\right) \sim (0, \sigma^2_{\epsilon_{p,tk}})$:

Error term composed of the following combined effects: interaction between Monte-Carlo and policy effects $(\mu_{tk}^{(p)} * \gamma_{itk})$ and the effects of previous biography outcomes $\underset{\sim}{\beta}_{i,t-1}$ and $\underset{\sim}{\beta}_{it[1:k-1]}$ on $F_{D_{tk}}(\cdot)$.

The statistical representation for $y_{itk}^{(p)}(r_{itk})$ developed above may be easily extended to obtain a time-aggregate mean model for the kth characteristic (e.g., medical cost) of biography i. Defining $\bar{\mu}_{.k}^{(p)} \equiv \frac{1}{T_i}\sum_{t=1}^{T_i} \mu_{tk}$, $\bar{\gamma}_{i.k} \equiv \frac{1}{T_i}\sum_{t=1}^{T_i} \gamma_{itk}$, and $\bar{\epsilon}_{i.k}^{(p)} \equiv \frac{1}{T_i}\sum_{t=1}^{T_i} \epsilon_{itk}^{(p)}$ yields the following linear model:

$$\bar{y}_{i.k}^{(p)} = \bar{\mu}_{.k}^{(p)} + \bar{\gamma}_{i.k} + \bar{\epsilon}_{i.k}^{(p)} \tag{5.5}$$

where

$\bar{\mu}_{.k}^{(p)}$ is the time-aggregate mean for characteristic k under policy p for the simulated population;

$\bar{\gamma}_{i.k}^{(p)} \sim (0, \sigma^2_{\gamma_k})$ is the time-aggregate mean Monte-Carlo effect mean induced by the selection of randomly drawn numbers $(r_{i1k}, r_{i2k}, \ldots, r_{iT_ik})$.

$\overline{\epsilon}_{i.k}^{(p)} \sim (0,\ \sigma_{\epsilon_k}^2)$ is the time-aggregate mean error component.

5.2 The D1 (conventional) and the D2 (proposed) Biography Generation Designs

The statistical representation for MSM developed in Section 5.1 may be used to estimate policy effects under various microsimulation designs. The statistical model may also be used to demonstrate the statistical gains of the D2 biography generation design proposed in Section 4.5. First, however, it is necessary to specify the type of alternative design (D1 design) we will be comparing the proposed design with.

In Section 4.5 the optimal D2 design was proposed to block out across-policy Monte-Carlo effects induced by differences in the random numbers ($r_{itk} \neq s_{itk}$) used to simulate the common decision steps for each biography across policy runs. The D2 design ensures that the same random numbers are used to simulate the common FDS segment of decision steps among all possible decision paths in $U_t = \{\underset{\sim}{D}_{t1}, \ldots, \underset{\sim}{D}_{tQ_t}\}$ for all times $t = 1, \ldots, T$. Under this design $r_{itk} = s_{itk}$ for $\forall_{i.t.k}$. The D2 design may be conveniently implemented using the Minimum Life Cycle biography generation scheme described in the Appendix.

We will refer to a biography generation design which does not consiously ensure that the same Monte-Carlo draws are used to simulate each common decision step across policy runs for all biographies as the D1 design. In this design, r_{itk} and s_{itk} are drawn independently of each other so that $\mathrm{Cov}(r_{itk}, s_{itk}) = 0$, $\forall_{itk}$. This further implies that the random variates r_{itk} and s_{itk} used to execute each decision step D_{tk} under policy $\alpha^{(1)}$ and $\alpha^{(2)}$, respectively, will be different in most cases: $r_{itk} \neq s_{itk}, \forall_{itk}$.

Frequently, when studying various policies the experimenter may attempt to intuitively reduce Monte-Carlo variation among simulated outcomes $y_{itk}^{(p)}$ obtained under different policy scenarios by storing the first initial seed r_{110} (i=1, t=1) used to start the simulation run under the first, say baseline, policy. Next, when performing simulations under other policy scenarios, the same initial seed is used to start the run in an attempt to reduce Monte-Carlo variation across policies.

This design maintains random number consistency (with decision steps) across different policy scenarios for all simulated biographies only under very stringent and usually untenable conditions. Not only must the state space exhibit the FDS (Fixed Decision Sequence) structure (discussed in Section 3.2) with only one possible decision sequence $\underset{\sim}{U}_t = \left\{\left(\mathrm{D}_{t1}, \ldots, \mathrm{D}_{ts_t}\right)\right\}$ arising at all times, but, additionally, the termination time must be the same for all biographies at all times: $T_i = T$, $i = 1, .., n$.

If the above conditions prevail then regardless of what policy scenario is in effect, the total number of Monte-Carlo draws ($\sum_{i=1}^{T_i} S_t$) needed to simulate each biography will be constant. In most MSM models even if the PDS condition holds for all simulation times $t = 1, .., \omega$, the condition $T_i = T$ for $i = 1, .., n$ is too limiting because the termination time is usually itself a random variable of interest.

Therefore, exempting the first simulated individual $i=1$, this design will fail to maintain random number consistency for all simulated biographies and is not sufficient to ensure that the same individual - in the Monte-Carlo "genetic" sense - is simulated under different policies. Hence, simply using the same starting seed fails to maintain decision step/random number consitency across policy scenarios and will not block out across-policy Monte-Carlo effects.

In their paper on "simulation experiments", Schruben and Margolin (1978) pay attention only to control over the use of the random number stream at different design "points". The issue of consistency between the decision path and the Monte-Carlo stream drawn to simulate it is ignored. Based on the discussion above, this approach will block out counfounding across-policy Monte-Carlo effects only in the most simple longitudinal microsimulation models.

5.3 Policy Comparison Methodology under the D1 Biography Generation Design

Without loss of generality, the D1 and D2 designs will be examined under the framework of two policy scenarios: $\alpha^{(1)}$ and $\alpha^{(2)}$. We further restrict our analysis to continuous (non-categorical) simulation outcomes and will assess the impact of policy differences by comparing simulation means using the *t*-test (ANOVA) approach. In cases where binary or categorical outcomes are generated, other statistical model (eg. logistic, multinominal) representations for MSM output may be specified and the basic concepts developed in the paper may be easily applied to these settings.

Let $y_{itk}^{(1)}$ and $y_{itk}^{(2)}$ denote respectively the simulated response for decision step D_{tk} under policy $\alpha^{(1)}$ and policy $\alpha^{(2)}$ for biography i at time t. To perform the *t*-test analysis, the means $\bar{y}_{tk}^{(1)}$ and $\bar{y}_{tk}^{(2)}$ are formed and (under assumption of independence of the two samples and normality) are used to test the null hypothesis that the means for the two policies $\mu_{tk}^{(1)}$ and $\mu_{tk}^{(2)}$ are equal: H_0: $\mu_{tk}^{(1)} - \mu_{tk}^{(2)} = 0$ where $\mu_{tk}^{(1)} - \mu_{tk}^{(2)}$ is estimated by $\bar{\Delta}_{tk} = \bar{y}_{tk}^{(1)} - y_{tk}^{(2)}$.

Our posited statistical model for an observation from MSM under this design is the same as the general model of equation (5.8):

$$y_{itk}^{(p)}(r_{itk}) = \mu_{tk}^{(p)} + \gamma_{itk} + \epsilon_{itk}^{(p)} \quad \begin{matrix} p = 1, 2 \\ i = 1, \ldots, n \\ t = 1, \ldots, T_i \\ k = 1, \ldots, S_t. \end{matrix} \tag{5.8}$$

with the individual terms assumed to exhibit the following standard behavioral assumptions:

$$E(\gamma_{itk}) = 0,\ E(\epsilon_{itk}^{(p)}) = 0,\ \mathrm{Var}\,(\gamma_{itk}) = \sigma^2_{\gamma_{tk}},\ \mathrm{Var}\,(\epsilon_{itk}^{(p)}) = \sigma^2_{\epsilon_{p,tk}},\ \mathrm{Cov}\,(\gamma_{itk},\ \epsilon_{itk}^{(p)}) = 0.$$

In the D1 design the random variates r_{itk} and s_{itk} used to respectively generate outcomes $y_{itk}^{(p)}(\cdot)$ under the two policy scenarios $\alpha^{(p)}, p = 1,2$, are drawn independently and will be usually different $(r_{itk} \neq s_{itk})$. Since $\mathrm{Cov}\,(r_{itk}, s_{itk}) = 0,\ \forall_{itk}$, then $\mathrm{Cov}\left(y_{itk}^{(1)}(r_{itk}),\ y_{itk}^{(2)}(s_{itk})\right) = 0,\ \forall_{itk}$. As a consequence, simulated outcomes $y_{itk}^{(p)}$, $p = 1,2$, for the same "genetic" individual under policies $\underset{\sim}{\alpha}_t^{(p)}$, $p = 1,2$, cannot be generated. Hence, we can treat the output from the two policy simulations as two independent samples.

Under the D1 design the mean difference $\bar{\Delta}_{tk}$, and its variance under policy p=1 and p=2 may be expressed as

$$\begin{aligned}\bar{\Delta}_{tk} &\overset{(D1)}{=} \bar{y}_{tk}^{(1)} - \bar{y}_{tk}^{(2)} \\ &= (\mu_{tk}^{(1)} - \mu_{tk}^{(2)}) + \frac{1}{n}\left(\sum_{i=1}^{n} \gamma_{itk} - \sum_{i=1}^{n} \gamma'_{itk}\right) + \frac{1}{n}\left(\sum_{i=1}^{n} \epsilon_{itk}^{(1)} - \sum_{i=1}^{n} \epsilon_{itk}^{(2)}\right).\end{aligned} \tag{5.9}$$

$$\mathrm{Var}\,(\bar{\Delta}_{tk}) \overset{(D1)}{=} (2\,\sigma^2_{\gamma_{tk}} + \sigma^2_{\epsilon_{1,tk}} + \sigma^2_{\epsilon_{2,tk}}) \,/\, n = \sigma^2_{D1}/n. \tag{5.10}$$

The corresponding t-test statistic under H_o is

$$\frac{\bar{y}_{tk}^{(1)} - \bar{y}_{tk}^{(2)}}{\sqrt{\mathrm{Var}\,(\bar{y}_{tk}^{(1)} - \bar{y}_{tk}^{(2)})}} = \frac{\bar{y}_{tk}^{(1)} - \bar{y}_{tk}^{(2)}}{\hat{\sigma}_{D1}/\sqrt{n}} \sim t_{2(n-1)}$$

where

$$\mathrm{Var}_{D1}(\bar{y}_{tk}^{(1)} - \bar{y}_{tk}^{(2)}) = \hat{\sigma}^2_{D1}/n = \frac{1}{n}\sum_{p=1}^{2}\sum_{i=1}^{n} \frac{(y_{itk}^{(p)} - \bar{y}_{tk}^{(p)})^2}{2(n-1)} \tag{5.11}$$

is an unbiased estimator of Var $(\bar{y}_{tk}^{(1)} - \bar{y}_{tk}^{(2)}) = (2\,\sigma^2_{\gamma_{tk}} + \sigma^2_{\epsilon_{1,tk}} + \sigma^2_{\epsilon_{2,tk}})/n$ based on the total within policy sum of squares error.

Additionally, when the impact of more than two policies needs to be compared, the ANOVA approach may be used. The relevant ANOVA table needed to carry out this analysis is given in Table 1 below (ignoring subscript tk):

Table 1: ANOVA TABLE under D1 Biography Generation Designs

EFFECT	SS	d.f.	E(MSS)	F
POLICY	$SS_P = \sum_p^P \sum_i^n (\bar{y}^{(p)} - \bar{y})^2$	$(P-1)$	$n\sum_p^P (\mu^{(p)})^2 + \sigma_\gamma^2 + \frac{\sum_p^P \sigma_{\epsilon(p)}^2}{P}$	$\frac{MSS_{POL}}{MSS_{MC}}$
Monte-Carlo (Within)	$SS_{MC} = \sum_p \sum_i (y_i^{(p)} - \bar{y}^{(p)})^2$	$P(n-1)$	$\sigma_\gamma^2 + \sum_p \frac{\sigma_{\epsilon(p)}^2}{P}$	
Total	$\sum_p \sum_i (y_i^{(p)} - \bar{y})^2$	$nP-1$		

From the expected mean sum of squares column it can be seen that the effect of policy may be tested (H_0: $\delta^{(1)} = \delta^{(2)} = .. \delta^{(p)}$) by using the F-test based on the ratio $\frac{MSS_{POL}}{MSS_{MC}} \sim F_{P-1,(n-1)P}$.

Can we do better? More precisely (going back to the two policy context), this is to ask if it is possible to obtain an unbiased estimator for the policy effect $\mu_{tk}^{(1)} - \mu_{tk}^{(2)}$ which is more efficient. It is shown in the next section that the answer to this question is yes - when the D2 biography generation design is used.

5.4 Policy Comparison Methodology under the D2 Biography Generation Design

Under the D2 design, the statistical model for MSM output

$$y_{itk}^{(p)}(r_{itk}) = \mu_{tk}^{(p)} + \gamma_{itk} + \epsilon_{itk}^{(p)} \qquad \begin{matrix} p = 1, 2 \\ i = 1, \ldots, n \\ t = 1, \ldots, T_i \\ k = 1, \ldots, S_t. \end{matrix}$$

remains the same. However, contrary to what occurred under the D1 design, simulated outputs $y_{itk}^{(p)}$, $p = 1, 2$, for the same "genetic" individual i are generated under policies $\underset{\sim}{\alpha}_t^{(p)}$, $p = 1, 2$ by using the same random draws r_{itk}, $\forall_{i,t,k}$ for all common decision steps in the FDS segment. This has the important consequence that

$$\mathrm{Cov}\left(y_{itk}^{(1)}(r_{itk}), y_{itk}^{(2)}(r_{itk})\right) = \mathrm{Cov}(\gamma_{itk}, \gamma_{itk}) = \sigma_{\gamma_{tk}}^2. \tag{5.12}$$

The mean difference $\bar{\Delta}_{tk}$ under policies p=1 and p=2 now becomes

$$\begin{aligned} \bar{\Delta}_{tk} &\overset{(D2)}{=} \bar{y}_{tk}^{(1)} - \bar{y}_{tk}^{(2)} \\ &= (\mu_{tk}^{(1)} - \mu_{tk}^{(2)}) + \frac{1}{n}\left(\sum_{i=1}^n \epsilon_{itk}^{(1)} - \sum_{i=1}^n \epsilon_{itk}^{(2)}\right) \end{aligned} \tag{5.13}$$

with variance

$$\mathrm{Var}(\bar{\Delta}_{tk}) \overset{(D2)}{=} (\sigma_{\epsilon_{1,tk}}^2 + \sigma_{\epsilon_{2,tk}}^2) / n = \sigma_{D2}^2 / n. \tag{5.14}$$

An unbiased estimator of σ^2_{D2} under this design is given by $\hat{\sigma}^2_{D2} = \frac{1}{n-1}\sum_{i=1}^{n}(\Delta_{itk} - \bar{\Delta}_{tk})^2$ as it can be easily shown that $E(\hat{\sigma}^2_{tk}) = \sigma^2_{\epsilon_{1,tk}} + \sigma^2_{\epsilon_{1,tk}} = \sigma^2_{D2}$.

Moreover, an unbiased estimator of $\mathrm{Var}(\bar{\Delta}_{tk})$ can now be constructed as

$$\hat{\mathrm{Var}}(\bar{\Delta}_{tk}) = \frac{\hat{\sigma}^2_{D2}}{n} = \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(\Delta_{itk} - \bar{\Delta}_{tk})^2. \tag{5.15}$$

In order to test for policy differences $(H_0: \mu^{(1)}_{tk} - \mu^{(2)}_{tk} = 0)$ we use the fact that

$$\frac{\bar{\Delta}_{tk}}{\sqrt{\hat{\mathrm{Var}}(\bar{\Delta}_{tk})}} = \frac{\bar{y}^{(1)}_{tk} - \bar{y}^{(2)}_{tk}}{\hat{\sigma}_{D2}/\sqrt{n}} \sim t_{n-1}$$

is distributed as a t-statistic with $n-1$ degrees of freedom.

Hence, the D2 design blocks out Monte-Carlo variability in the estimator for policy difference $\bar{\Delta}_{tk} = \bar{y}^{(1)}_{tk} - \bar{y}^{(2)}_{tk}$, leading to the use of a paired t-test methodology in comparing policy differences.

5.5 D1 and D2 Designs: Efficiency and Stability

It may be seen that the D2 design leading to the use of the paired t-test is a superior strategy over the combined D1 design and unpaired t-test (ANOVA) strategy in two important respects:

i) In the context of comparing simulation means to assess the impact of policy change the estimator of policy differences $\bar{\Delta}_{tk} = \bar{y}^{(1)}_{tk} - \bar{y}^{(2)}_{tk}$ is more efficient under the D2 design:

$$\mathrm{Var}_{D1}(\bar{\Delta}_{tk}) - \mathrm{Var}_{D2}(\bar{\Delta}_{tk}) = 2\,\sigma^2_{\gamma_{tk}}/n > 0.$$

This occurs because the D2 biography generation design blocks out the component of Monte-Carlo variability due to γ_{itk} in the difference of individual outcomes $\Delta_{itk} = y^{(1)}_{itk} - y^{(2)}_{itk}, \forall_i$, by ensuring that in the Monte-Carlo sense the same "genetic" individual is observed under both policies.

ii) The estimator $\bar{\Delta}_{tk} = \bar{y}^{(1)}_{tk} - \bar{y}^{(2)}_{tk}$ of the policy difference under the D2 design is more stable in smaller samples. This is implied by i) but also by the terms in $\bar{\Delta}$ as shown by expressions (5.9) and (5.13) respectively under the D2 design:

$$\bar{\Delta}_{tk} \overset{D1}{=} (\mu_{tk}^{(1)} - \mu_{tk}^{(2)}) + \frac{1}{n}(\sum_{i=1}^{n} \gamma_{itk} - \sum_{i=1}^{n} \gamma'_{itk}) + \frac{1}{n}(\sum_{i=1}^{n} \epsilon_{itk}^{(1)} - \sum_{i=1}^{n} \epsilon_{itk}^{(2)}) \tag{5.9}$$

$$\bar{\Delta}_{tk} \overset{D2}{=} (\mu_{tk}^{(1)} - \mu_{tk}^{(2)}) + \frac{1}{n}(\sum_{i=1}^{n} \epsilon_{itk}^{(1)} - \sum_{i=1}^{n} \epsilon_{itk}^{(2)}) \tag{5.13}$$

Note that $\bar{\Delta}_{tk}$ under both designs is consistent, however, in small samples the term $(\sum_{i=1}^{n} \gamma_{itk} - \sum_{i=1}^{n} \gamma'_{itk})/n$ may not be exactly zero. However, this term is completely eliminated by the D2 design leading to more stable behavior in small samples.

5.6 Statistical Gains from the D2 Design: Significance Level, Power, and Sample Size

To conclude this discussion on the statistical gains of using the D2 biography generation design, we consider the implications of the fact that paired differences for data generated over policies $\underset{\sim}{\alpha}_t^{(p)}$, $p = 1, 2$, under the D2 design yields smaller variances than the D1 design: $\sigma_{D2}^2 \leq \sigma_{D1}^2$.

Without loss of generality, we discuss the import of the above result within the context of the simple hypothesis H_0: $\mu_{tk}^{(1)} - \mu_{tk}^{(2)} = \mu_0 = 0$ versus H_1: $\mu_{tk}^{(1)} - \mu_{tk}^{(2)} = \mu_1 = > 0$ under the assumption of normality with σ^2 known. The corresponding equation for type I error is given by (droping the subscript tk)

$$\alpha = Pr\{\frac{\bar{\Delta} - \mu_o}{\sigma/\sqrt{n}} \geq Z_\alpha\}$$

where

$$Z_\alpha = \frac{(k_\alpha - \mu_o)}{\sigma}\sqrt{n} \tag{5.16}$$

and k_α is the critical value for $\bar{\Delta}$ consistent with a test of significance level α.

Similarly, the power equation is

$$1 - \beta = Pr\{\frac{\bar{\Delta} - \mu_1}{\sigma/\sqrt{n}} \geq Z_{1-\beta}\}$$

where

$$Z_{1-\beta} = \frac{(k_\alpha - \mu_1)}{\sigma}\sqrt{n}. \tag{5.17}$$

Combining (5.16) and (5.17) together yields

$$Z_{1-\beta} = Z_\alpha - (\mu_1 - \mu_o)\frac{\sqrt{n}}{\sigma} \tag{5.18}$$

Expression (5.18) allows us to evaluate the impact of the D2 design against the D1 design when $\sigma^2_{D2} \leq \sigma^2_{D1}$ with respect to type I error, power, and sample size. Some interesting cases of spreading efficiency gains are considered below.

i) Effect on Sample Size with Constant Significance Level and Power

In this scenario $\alpha_{D1} = \alpha_{D2}$ and $(1 - \beta_{D1}) = (1 - \beta_{D2})$. Applying equation (5.18) to the D1 and D2 situations yields

$$D1: \quad Z_{1-\beta} = Z_{\alpha} - (\mu_1 - \mu_o)\frac{\sqrt{n_{D1}}}{\sigma_{D1}} \tag{5.19}$$

$$D2: \quad Z_{1-\beta} = Z_{\alpha} - (\mu_1 - \mu_o)\frac{\sqrt{n_{D2}}}{\sigma_{D2}} \tag{5.20}$$

These two equations reveal that $n_{D2} = \frac{\sigma^2_{D2}}{\sigma^2_{D1}} n_{D1} \leq n_{D1}$. (5.21)

Result (5.21) implies that the same significance level and power offered by the D1 design may be attained at a lower sample size n_{D2} ($\leq n_{D1}$) with the D2 design. The extent of the reduction in sample design $n_{D1} - n_{D2}$ depends on the relative magnitudes of σ^2_{D1} and σ^2_{D2}. More precisely, since σ^2_{D1} and σ^2_{D2} are related as (see equations 5.9 and 5.13)

$$\sigma^2_{D1} = 2\,\sigma^2_{\gamma_{tk}} + \sigma^2_{D2} \ ,$$

the reduction in sample size (droping the subscripts for time t and decision k) is exactly

$$n_{D1} - n_{D2} = \frac{2\,\sigma^2_{\gamma}}{\sigma^2_{\epsilon_1} + \sigma^2_{\epsilon_2} + 2\,\sigma^2_{\gamma}} n_{D1}.$$

Recall that $2\,\sigma^2_{\gamma}$ is the additional across-policy Monte-Carlo variability induced by using different Monte-Carlo draws in executing decision D_{tk} under policies $\underset{\sim}{\alpha}_i^{(p)}$, $p = 1, 2$, when generating outputs $\underset{\sim}{y}_{ik}^{(p)}$, $p = 1, 2$. . Under the D2 design, however, the term $2\,\sigma^2_{\gamma}$ is totally eliminated, reducing with it the sample size by the amount $\frac{2\,\sigma^2_{\gamma}}{\sigma^2_{\epsilon_1} + \sigma^2_{\epsilon_2} + 2\,\sigma^2_{\gamma}} n_{D1}$ while maintaining the same level of significance and power for the testing procedure.

ii) Effect on Significance Level with Constant Sample Size and Power

In this scenario $n_{D1} = n_{D2}$ and $(1 - \beta_{D1}) = (1 - \beta_{D2})$. Applying equation (5.18) to this situation yields

$$Z_{\alpha_{D2}} = Z_{\alpha_{D1}} + (\mu_1 - \mu_o)\frac{(\sigma_{D1} - \sigma_{D2})}{\sigma_{D1}\,\sigma_{D2}}\sqrt{n}\,. \tag{5.22}$$

Since $\mu_1 > \mu_0$ and $\sigma_{D1} > \sigma_{D2}$, this implies that $Z_{\alpha_{D2}} > Z_{\alpha_{D1}}$ so that $\alpha_{D1} > \alpha_{D2}$. Therefore, one implication of the D2 design is that when the sample size and power achieved under the D1 design are held constant, a smaller type I error (higher significance level) results for the test. The magnitude of the gain in the significance level is quantified by equation (5.18) where Z_α is determined by the distibution of the generated data.

If $\Phi_\alpha(\cdot)$ is allowed to represent the cdf of the transformed random variable $\dfrac{\overline{\Delta}_{tk} - \mu_0}{S.E.(\overline{\Delta}_{tk})}$, then the reduction in type I error may be further quantified as

$$\Delta\alpha = \alpha_{D1} - \alpha_{D2} = \Phi(Z_{\alpha_{D2}}) - \Phi(Z_{\alpha_{D1}}) > 0. \tag{5.23}$$

iii) Effect on Power with Constant Sample Size and Significance Level

In this case $n_{D1} = n_{D2}$ and $\alpha_{D1} = \alpha_{D2}$ and use of equation (5.18) leads to

$$Z_{1-\beta_{D2}} = Z_{1-\beta_{D1}} - (\mu_1 - \mu_o)\frac{(\sigma_{D1} - \sigma_{D2})}{\sigma_{D1}\,\sigma_{D2}}\sqrt{n}\,. \tag{5.24}$$

Since $\mu_1 > \mu_0$ and $\sigma_{D1} > \sigma_{D2}$, $Z_{1-\beta_{D2}} < Z_{1-\beta_{D1}}$ so that $(1-\beta_{D1}) < (1-\beta_{D2})$. Hence, when the type I error and the sample size are held fixed, all gains from the D2 design are tranferred to achieving a higher level of power: $(1-\beta_{D1}) < (1-\beta_{D2})$. Similarly, with $\Phi(\cdot)$ defined as the cdf of the transformed random variable $\dfrac{\overline{\Delta}_{tk} - \mu_1}{S.E.(\overline{\Delta}_{tk})}$, the increase in power may be quantified as

$$\Delta(1-\beta) = (1-\beta_{D2}) - (1-\beta_{D1}) = \Phi(Z_{1-\beta_{D1}}) - \Phi(Z_{1-\beta_{D2}}) > 0. \tag{5.25}$$

6. SIMULATION STUDY

A simulation study was performed to empirically verify the theoretical results regarding the D2 biography generation design obtained in earlier sections within the context of a simple microsimulation model. The microsimulation model used consists of a FDS sequence of two decision steps D_1 and D_2 with only one time point ($T=1$). Decision step D_1 simulates the smoking status of the biography while D_2 simulates the future lifetime of the biography. In terms of the notation and concepts developed in the paper, the complete microsimulation model is given in Table 2 below.

Table 2. Microsimulation Model Used in Simulation Study

	D_1	D_2
Decision Sequence		
$\underset{\sim}{D} = (D_1, D_2)$	(smoking status)	(time to death)
Decision Space		
X_k	$x_1 = \{NS, LS, MS, HS\}$	$x_2 = [0, \infty)$
Transition pdf		
$f_{D_k \mid D_{k-1}}(\beta_{ik}; \underset{\sim}{\theta}_{ik}^{(p)}(\underset{\sim}{\alpha}^{(p)}, \underset{\sim}{\beta}_{i,k-1})$		
i) Policy $\underset{\sim}{\alpha}^{(1)} = (\underset{\sim}{\alpha}_1^{(1)}, \underset{\sim}{\alpha}_2^{(1)})$		$f_{D_2/D_1}(\beta_{i2}; \theta_{i2}^{(1)}(\underset{\sim}{\alpha}_2^{(1)}, \beta_{i1})$ ~ Weibul $(\theta_{i2}^{(1)}, C=3)$
	$f_{D_1}(\beta_{i1}; \underset{\sim}{\theta}_{i1}^{(1)}(\underset{\sim}{\alpha}_1^{(1)})) = \begin{cases} .3 \text{ if } \beta_{i1} = NS \\ .2 \text{ if } \beta_{i1} = LS \\ .3 \text{ if } \beta_{i1} = MS \\ .2 \text{ if } \beta_{i1} = HS \end{cases}$	$\underset{\sim}{\theta}_{i2}^{(1)}(\underset{\sim}{\alpha}_2^{(1)}, \beta_{i1}) = \begin{cases} 1/60 \text{ if } \beta_{i1} = NS \\ 1/58 \text{ if } \beta_{i1} = LS \\ 1/56 \text{ if } \beta_{i1} = MS \\ 1/54 \text{ if } \beta_{i1} = HS \end{cases}$
ii) Policy $\underset{\sim}{\alpha}^{(2)} = (\underset{\sim}{\alpha}_1^{(2)}, \underset{\sim}{\alpha}_2^{(2)})$		$f_{D_2/D_1}(\beta_{i2}; \theta_{i2}^{(2)}(\underset{\sim}{\alpha}_2^{(2)}, \beta_{i1})$ ~ Weibul $(\theta_{i2}^{(2)}, C=3)$
	$f_{D_1}(\beta_{i1}; \underset{\sim}{\theta}_{i1}^{(2)}(\underset{\sim}{\alpha}_1^{(2)})) = \begin{cases} .6 \text{ if } \beta_{i1} = NS \\ .2 \text{ if } \beta_{i1} = LS \\ .1 \text{ if } \beta_{i1} = MS \\ .1 \text{ if } \beta_{i1} = HS \end{cases}$	Since $\underset{\sim}{\alpha}_2^{(1)} = \underset{\sim}{\alpha}_2^{(2)}$, $\theta_{i2}^{(1)} = \theta_{i2}^{(2)}$

Note that the policy intervention ($\underset{\sim}{\alpha}^{(2)}$) changes only the distribution of smokers, with a larger segment of the population shifting towards lower smoking. The policy change does not effect the parameter $\theta_{i2}^{(p)}$ to $f_{D_2 \mid D_1}(\cdot)$ as $\underset{\sim}{\alpha}_2^{(1)} = \underset{\sim}{\alpha}_2^{(2)}$. Moreover, the pdf supporting the decision space X_2 for decision step D_2 may be written analytically as follows:

$$f_{D_2 \mid D_1}(\beta_{i2}; \theta_{i2}, C=3) = \frac{1}{C\theta_{i2}} [\exp(-\beta_{i2}/\theta_{i2})]^{1/c}$$

where the shape parameter C is fixed at the value C = 3. The value of θ_2 faced by the individual biography under policy $\underset{\sim}{\alpha}^{(p)}$ represented by $\theta_{i2}^{(p)}(\underset{\sim}{\alpha}_2^{(p)}, \beta_{i1})$ depends on the smoking status (β_{i1}) generated for the biography in performing the first decision step D_1.

To be consistent with the notation used in the paper in Sections 5 we define $y_{ik}^{(p)} (\equiv \beta_{ik})$ to be the simulation output of analytical interest for biography i for decision step D_k under policy $\underset{\sim}{\alpha}^{(p)}$. The expected theoretical difference for the future life-time random variable $y_{i2}^{(p)}$ under policy scenarios $\underset{\sim}{\alpha}^{(1)}$ and $\underset{\sim}{\alpha}^{(2)}$ is $E(y_{i2}^{(1)}) - E(y_{i2}^{(2)}) = -1.2502$ years. This means that if the "smoking distribution" of the population is perturbed by health/education programs to effectuate $\underset{\sim}{\alpha}^{(2)}$, the expected future life-time should increase by 1.2502 years over the "smoking distribution" reflected by current policy $\underset{\sim}{\alpha}^{(1)}$.

A program was written in GAUSS to simulate the microsimulation model described above. One hundred ($K = 100$) simulations were performed under varying sample sizes in the range from 5 to 3,000 biographies under both the D_1 and D_2 biography generation designs. For each simulation the means $\bar{y}_2^{(1)}$ and $\bar{y}_2^{(2)}$ for the observed future life-times were computed under the two policy scenarios along with the difference in means $\bar{\Delta} = \bar{y}_2^{(1)} - \bar{y}_2^{(2)}$ and an estimate of its variance $\hat{\mathrm{Var}}(\bar{\Delta})$ under the D_1 and D_2 designs using the expressions obtained in Sections 5.3 and 5.4.

Table 3. Variance of $\bar{\Delta}$ and its Estimators

	D1 Design	**D2 Design**
$\mathrm{Var}(\bar{\Delta})$	σ_{D1}^2/n	σ_{D2}^2/n
σ^2	$\sigma_{D1}^2 = (2\sigma_\gamma^2 + \sigma_{\epsilon_2^{(1)}}^2 + \sigma_{\epsilon_2^{(2)}}^2)$	$\sigma_{D2}^2 = (\sigma_{\epsilon_2^{(1)}}^2 + \sigma_{\epsilon_2^{(2)}}^2)$
$\hat{\mathrm{Var}}(\bar{\Delta})$	$\hat{\sigma}_{D1}^2/n$	$\hat{\sigma}_{D2}^2/n$
$\hat{\sigma}^2$	$\hat{\sigma}_{D1}^2 = \frac{1}{2(n-1)} \sum_{p=1}^{2} \sum_{i=1}^{n} (y_i^{(p)} - \bar{y}^{(p)})^2$	$\hat{\sigma}_{D2}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \sum (\Delta_i - \bar{\Delta})^2$ where $\Delta_i = y_{i2}^{(1)} - y_{i2}^{(2)}$

Estimates of the policy effect $\mu_\Delta = \mu_2^{(1)} - \mu_2^{(2)}$ and its standard error σ obtained under the D1 and D2 biography generation designs under varying sample sizes averaged over the $k = 100$ simulations are reported in Table 4 below.

Table 4. Results from Simulation Study: Averages of $\bar{\Delta}$ and $\hat{\sigma}$ Under D1 and D2 Designs over 100 Simulations

	D1 Design		D2 Design		
n	$\bar{\Delta}_{D1}$ (-1.250)	$\bar{\hat{\sigma}}_{D1}$	$\bar{\Delta}_{D2}$ (-1.250)	$\bar{\hat{\sigma}}_{D2}$	$\bar{\hat{\sigma}}^2_{D1} / \bar{\hat{\sigma}}^2_{D2}$
5	.895	19.413	-1.326	1.258	238
10	-1.015	19.218	-1.243	1.271	229
15	-.343	18.903	-1.270	1.275	231
20	-2.355	19.120	-1.235	1.258	220
50	-1.7192	18.747	-1.258	1.272	218
100	-1.490	18.843	-1.254	1.298	212
200	-1.282	18.818	-1.253	1.294	211
500	-1.203	18.970	-1.250	1.304	212
1000	-1.221	18.926	-1.253	1.300	212
2000	-1.298	18.836	-1.251	1.298	213
3000	-1.255	18.884	-1.252	1.299	211

Comparing the columns for $\bar{\hat{\sigma}}$ under the D1 and D2 designs along with the column for $\bar{\hat{\sigma}}^2_{D1} / \bar{\hat{\sigma}}^2_{D2}$ reveals the extraordinary efficiency of the D2 design over the D1 design (approximately of the order 211 times). In Section 5.6, the relationship between the sample sizes n_{D2} and n_{D2} under the two designs required maintaining the same type I error ($\alpha_{D1} = \alpha_{D2}$) and level of power $(1-\beta_{D1}) = (1-\beta_{D2})$ was given by

$$n_{D1} = \frac{\sigma^2_{D1}}{\sigma^2_{D2}} n_{D2}.$$

Using this result with our observed estimate of $\hat{\sigma}^2_{D_1} / \hat{\sigma}^2_{D_2} \simeq 211$ in this particular simulation experiment, one interpretation of the extremely high efficiency of the D2 design is that a sample size approximately 211 times larger is required for the D1 design to maintain the same significance level and power for the testing procedure as fixed for the D2 design.

It should be kept in mind that the ratio $\sigma^2_{D1} / \sigma^2_{D2} = \left(2\sigma^2_{\gamma_2} / \sigma^2_{\epsilon_2^{(1)}} + \sigma^2_{\epsilon_2^{(2)}}\right) + 1$ is large because $2\sigma^2_{\gamma_2} >> \sigma^2_{\epsilon_2^{(1)}} + \sigma^2_{\epsilon_2^{(2)}}$. In the simple microsimulation model used in this example $\sigma^2_{\epsilon_2^{(1)}} + \sigma^2_{\epsilon_2^{(2)}} = \mathrm{Var}(\epsilon^{(1)}_{i2}) + \mathrm{Var}(\epsilon^{(2)}_{i2})$ (the sum of the two error variances in our MSM model $y^{(p)}_{2i} = \mu^{(p)}_2 + \gamma_{i2} + \epsilon^{(p)}_{i2}$, $p = 1, 2$) is small. Recall that $\epsilon^{(p)}_{i2} \sim (0, \sigma^2_{\epsilon_2^{(p)}})$ representing the error component of

the model, is influenced by i) the interaction of policy and Monte-Carlo effects ($\underset{\sim}{\alpha}_t^{(p)} * \gamma_{itk}$) and by ii) perturbations to $\theta_{itk}^{(p)}$ ($\underset{\sim}{\alpha}_t^{(p)}, \underset{\sim}{\beta}_{i,t-1}, \underset{\sim}{\beta}_{it[1:k-1]}$), the parameter to the transition density $F_{D_{tk}}(\,;\theta_{itk}^{(p)})$, due to its dependence on previous outcomes $\underset{\sim}{\beta}_{i,t-1}$ and $\underset{\sim}{\beta}_{it[1:k-1]}$. In our simple model ($T = 1$ and $S_t = 2$), $\theta_{i2}^{(p)}(\alpha_t^{(p)}, \beta_{i1})$ is affected only by one previous outcome, namely β_{i1}, so that sources of perturbations to $\theta_{i2}^{(1)}$ as a result of variability in previous outcomes are smaller making $\sigma_{\epsilon_2^{(1)}}^2 + \sigma_{\epsilon_1^{(\infty)}}^2$ small in comparison to the Monte-Carlo variance $\mathrm{Var}(\gamma_{i2}) = \sigma_{\gamma_2}^2$.

In more complex microsimulation models we expect the term $\sigma_{D2}^2 = \sigma_{\epsilon_2^{(1)}}^2 + \sigma_{\epsilon_1^{(\infty)}}^2$ to be larger relative to $2\,\sigma_{\gamma_{tk}}^2$ so that the ratio $\dfrac{\sigma_{D1}^2}{\sigma_{D2}^2} = \dfrac{2\,\sigma_{\gamma_{tk}}^2}{\sigma_{\epsilon_2^{(1)}}^2 + \sigma_{\epsilon_1^{(\infty)}}^2} + 1$ will be smaller, although we expect it to be still much larger than one.

In Section 5.5 it was shown that the second advantage of the D2 design arose from the complete blocking out of the Monte-Carlo effect (γ_{itk}) in the estimator of policy differences $\bar{\Delta}_{tk}^{D2} = (\mu_{tk}^{(1)} - \mu_{tk}^{(2)}) + \dfrac{1}{n}\left(\sum_{i=1}^{n} \epsilon_{itk}^{(1)} - \sum_{i=1}^{n} \epsilon_{itk}^{(2)}\right)$ leading to its higher stability especially in small-scale simulations. This facet is also borne out by the simulation results by noting the closeness of $\bar{\Delta}_{D2}$ to the theoretical expected difference of -1.250 years even in small sample ages of $n = 5, 10, 15$. The estimator $\bar{\Delta}_{D1}$ under the D1 design behaves very poorly in these small samples and begins to display better consistency after exceeding $n = 100$.

7. CONCLUSION

In this paper, a theoretical description for the longitudinal microsimulation process was abstracted and used to identify an efficient biography generation design (D2 design). It was shown that policy changes inadvertently incur with them across-policy Monte-Carlo effects. To compensate, larger simulation experiments with greater number of simulated biographies are usually performed in the hope that i) across-policy Monte-Carlo perturbations are averaged out so that estimators of policy effects are stable and well behaved and ii) standard errors of policy effects decrease to yield the desired significance level and power for the testing procedure.

Using the mathematical representation, a statistical model for longitudinal microsimulation output was developed and used to demonstrate the high efficiency and stability of the estimator of policy effects under the D2 design. One implication of these gains in efficiency on the policy comparison methodology is that much smaller sample sizes are required to achieve a specified level of significance and power for the testing procedure.

In the simulation study performed using a small microsimulation model, it was found that the efficiency of the D2 design over the D1 design was approximately 211 times greater. This implies that under the D2 design, 211 times fewer biographies need to be generated over other designs which exercise no control over across-policy biography generation. This constitutes a significant saving in computing resources for very large scale simulations and also in statistical resources as less data needs to be tabulated and analyzed.

Finally, the advantage of the proposed D2 biography generation design is amplified when there is interest in the effect of policy change on rare events. Rare events may constitute an atypical condition observed in the population with very low frequency (eg. 30 cases per 100,000). In the study of rare-events, a certain minimum desired number of rare-event biographies need to be generated under different policy scenarios to allow statistically valid comparisons. The proposed biography generation design would reduce by a large factor the total number of biographies needed to measure and test the effects of policy change on the occurrence of rare events.

REFERENCES

Kleijnen, J.P.C. (1974). **Statistical Techniques in Simulation** (Part I). Marcel Dekker, Inc. New York

Kleijnen, J.P.C. (1979). **The Role of Statistical Methodology in Simulation. In Methodology in Systems Modelling and Simulation**. Zeigler, B.P. *et al* (Eds)

Kovacevic, M.C., Pandher, G.S, (1993). **Variance Estimation in Longitudinal Microsimulation**. Working Paper Series No. SSMD-93-010E, Methodology Branch, Statistics Canada.

Orcutt, G., Merz, J., Quinke,H. (Eds.) (1986). **Microanalytic Simulation Models to Support Social and Financial Policy**. North-Holland, Amsterdam.

Pawlikowski, K. (1990). **Steady-State Simulation of Queuing Processes: A Survey of Problems and Solutions**. ACM Computing Surveys, 22, 123-170.

Schmeiser, B. (1982). **Batch Size effects in the Analysis of Simulation Output**. Operations Research, 30, 556-568.

Schruben, L.W., Margolin, B.H. (1978). **Pseudorandom Number Assignment in Statistically Designed Simulation and Distribution Sampling Experiments**. Journal of the American Statistical Association, Vol. 73, 504-520.

Wolfson, M., Bertholet, J-M. (1992). **The Burden of CHD, Hypercholesteremia and Smoking: Exploratory Simulations Using POHEM**. Technical Report, Statistics Canada

Wolfson, M.C. (1992). **POHEM - A Framework for Understanding and Modelling the Health of Human Populations**. Research Paper Series No.44., Analytical Studies Branch, Statistics Canada.

APPENDIX: METHODS FOR IMPLEMENTING THE D2 BIOGRAPHY GENERATION DESIGN

The objective of the D2 biography generation design is to ensure that under all policy scenarios, each generated outcome $\beta_{it}^{(k)}\, i=1,\ldots,n;\ k=1,\ldots,S_t;\ T=1,\ldots,T_i$, simulated from decision steps in the common FDS segment of decision paths uses the same Monte-Carlo draws $r_{itk}, \forall_{itk}$ (see Section 4.4).

Random Number Generation in MSM

The Monte-Carlo draws in $\underset{\sim}{r}_{it} = (r_{it1}, \ldots, r_{itM_{it}})$, $M_{it} \le S_t$, used to simulate biography i at time t are generated iteratively from their preceding values. For instance, the first number in the random number stream of $\underset{\sim}{r}_{it}$ is generated from the seed r_{it0} and subsequent numbers are generated recursively using a relation of the form $r_{itk} = f(r_{it,k-1}),\ k \ge 1$. Due to the algorithmic dependence of the random variates used to simulate each biography on preceding values, the choice of the initial seed r_{i10} (t=1) effectively determines the sequence of random number streams $\underset{\sim}{R}_i = (\underset{\sim}{r}_{i1}, \ldots, \underset{\sim}{r}_{iT_i})$ used in simulating biography i till its time of death T_i. Upon completion of the i^{th} simulation, the last Monte-Carlo draw serves as the initial seed $r_{i+1,t0}$ for the next biography to be simulated. More precisely, $r_{i+1,10} = r_{iT_iS_{T_i}}$ where $r_{iT_iS_{T_i}}$ is the last random number to be used in $\underset{\sim}{r}_{iT_i} = (r_{iT_i1}, \ldots, r_{iT_iS_{T_i}})$ once the biography terminates at time $t=T_i$. Since the time to death T_i is random, the length of the Monte-Carlo stream $\underset{\sim}{R}_i$ used to simulate biography i is also random.

Methods for Maintaining Monte-Carlo Consistency Across Policies

Two methods for implementing the D2 design which ensure initial seeding consistency between policy runs are now described.

i) **Maximum Life Cycle Method**

Let M represent the maximum size of the random stream that may be required to simulate a biography through the longest path in the state space to reach the death state in the longest conceivable time T. We may further define M as the product of two components: $M=Tm$, where T represents the maximal life span of any biography and m is the upper limit on the maximal number of decision steps to appear in the longest decision path of $\underset{\sim}{U}_t = \{\underset{\sim}{D}_{t1}, \underset{\sim}{D}_{t2}, \ldots, \underset{\sim}{D}_{tQ_t}\}$ (the set of all possible decision paths at time t) over all times $t = 1,\ldots,T$. The D2 design may be implemented by simulating any biography i at time t by drawing the random vector $\underset{\sim}{r}_{it} = (r_{it1}, \ldots r_{itm})$ of fixed length m to generate $\underset{\sim}{\beta}_{it}$ from $\underset{\sim}{\beta}_{i,t-1}$. The complete Monte-Carlo stream drawn to generate biography i under this scheme may be expressed as

$\underset{\sim}{R}^{*} = (r_{i11,\ldots,} r_{i1m}, \; r_{i21}\cdots, r_{i2m}, \ldots, r_{iT1,\ldots,} r_{iTm})$.

We know from the definition of m that all transitions between states over all times will require at most m random variates. The surplus variates, those not required further during the transition, will not be used. Since M is known or determined beforehand, each biography will require a maximum of $Tm = M$ random numbers to complete its microsimulation. If the biography terminates before time T ($T_i < T$),, then the remaining random number stream in $\underset{\sim}{r}_i$ following r_{iT_im} will again remain un-used.

Under the Maximum Life Cycle scheme, upon completion of simulating biography i, the random number generator will advance to the last element of $\underset{\sim}{R}_i^{*}$, namely r_{iTm}, and this value will serve as the initial seed for the next $(i+1)$ simulation; i.e. $r_{i+1,0} = r_{iTm}$. This strategy of selecting initial seeds $r_{it0}, i = 1, \ldots, N, \; t = 1, \ldots, D_i$, will ensure that the corresponding random streams $\underset{\sim}{r}_{it}$ used to generate events in the FDS segment of decision paths will remain constant for all biographies under different policy scenarios (as long as the first seed is the same).

This method of drawing the random variates $\underset{\sim}{R}_i = \underset{\sim}{R}^{*} = (\underset{\sim}{r}_{i1}, \ldots, \underset{\sim}{r}_{iT})$, $i = 1, \ldots, n$, will maintain Monte-Carlo/Decision Step consistency across policies so that common simulated outcomes under different policy settings are generated without the confounding effects of across policy Monte-Carlo variation. This effectively means that, in the Monte-Carlo sense, one will observe the same "genetic" individuals under different policies.

ii) **Storing Initial Seeds**

In this approach, the initial random variates r_{it1}, $i=1,\ldots,n$, $t=1,\ldots,T_i$, used for each biography at each time under the first (baseline) policy $\underset{\sim}{\alpha}^{(1)}$ are stored to be re-used in the microsimulation of biographies under other policy scenarios $\underset{\sim}{\alpha}^{(p)}, p > 1$. Although this methods will also implement the D2 biography generation design of Section 4.4., its main drawback is that the memory required for storing initial Monte-Carlo draws is directly proportional to the size of the simulated population.