Proceedings of the 2020 Winter Simulation Conference
K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, eds.

DEVELOPING HIGH-QUALITY MICROSIMULATION MODELS USING R IN HEALTH DECISION SCIENCES

Heesun Eom
Yan Li

Department of Population Health Science & Policy
Icahn School of Medicine at Mount Sinai
One Gustave L. Levy Place, Box 1077
New York, NY 10029, USA

ABSTRACT

Health decision science is a growing field that studies the use of population health data and advanced analytical tools to inform decisions. This paper describes several modeling approaches and programming languages widely used in health decision sciences. Special emphasis is put on the development of microsimulation models using R. A recent microsimulation model—Simulation for Health Improvement and Equity (SHINE) Model—is described to demonstrate the development of microsimulaiton models using R. Several practical recommendations for developing microsimulation models using R are proposed. This paper may serve as a practical guide for population health scientists and healthcare professionals to develop their own microsimulation models to inform complex health decisions.

1 INTRODUCTION

Decision science is a broad field that uses existing data and analytical approaches to better inform decision-making (Hunink et al. 2014). The Center for Health Decision Science at Harvard T.H. Chan School of Public Health defines "decision science" as the following:

"While most fields of research focus on producing new knowledge, decision science is uniquely concerned with making optimal choices based on available information. Decision science seeks to make plain the scientific issues and value judgments underlying these decisions, and to identify tradeoffs that might accompany any particular action or inaction."

Decision science requires a highly interdisciplinary approach and often incorporates knowledge from a wide range of fields such as psychology, economics, mathematics, engineering, and statistics. Moreover, decision science can be applied in diverse fields such as business management, environmental regulations, disaster relief, and public health and policy. Any fields that deal with limited resources and quantifiable objectives can apply decision science to improve decision-making and increase effectiveness and efficiency.

This introductory tutorial paper will focus on health decision sciences which tackle research questions such as optimizing resource allocation for a health system, comparing the cost-effectiveness of different treatment strategies, assessing the health impacts of different public health policies (Hunink et al. 2014). Health decision sciences have utilized a wide array of quantitative approaches with varying levels of complexity. Simple approaches such as marginal analysis (i.e., weighing the benefit against the costs) can be useful for less sophisticated decisions. A more sophisticated approach such as decision tree allows decision-makers to deal with multiple variables at the same time. On the other end of the spectrum are sophisticated simulation-based decision science tools such as Markov models, microsimulations, and

agent-based models. These models can simulate health outcomes under different assumptions and policy scenarios.

In this paper, we describe several health decision science modeling approaches and programming languages in Section 2. Although every decision science approach and programming language has its pros and cons, microsimulation modeling and the R programming language are growing fast in health decision sciences over the past decade. We thus describe the development of microsimulation models using R in detail and provide a concrete example—a cardiovascular disease microsimulation model developed using R—to demonstrate the model development process in Section 3. Next, we provide a few recommendations for model development using R in Section 4. We discuss potential pitfalls of developing microsimulation models using R and point out future research directions in Section 5. At last, we provide brief concluding remarks in Section 6. This tutorial may be helpful for health scientists who are not familiar with microsimulation models or R to develop their first microsimulation model using R and promote the use of advanced simulation tools in health decision sciences.

2 OVERVIEW OF HEALTH DECISION-ANALYTIC MODELING

2.1 Decision Models

In this section, we will describe four decision-analytic modeling approaches often utilized in health decision sciences with varying levels of complexity, including decision trees, Markov models, microsimulations, and agent-based models. The first two modeling approaches are cohort-based approaches that focus on the modeling of population behaviors, while the other two are individual-based modeling approaches that focus on the modeling of health behaviors and outcomes at the individual level.

2.1.1 Decision Tree

Decision tree is a straightforward decision-making tool that can be developed even by those with little training in computer programming. As the name implies, decision tree models rely on a branching structure where a node represents splitting paths. Typically, the nodes are categorized into decision nodes, chance nodes, and end nodes and are represented by different symbols including squares, circles, and triangles, respectively, in a tree diagram. A health-related decision is made at the first node, the decision node. The probability of various outcomes is captured at the chance nodes. End nodes are absorbing nodes that usually represent death or final health outcomes (Kamiński et al. 2018).

Decision trees are easy to develop and understand and has thus become a useful decision-making tool for many health researchers. Analyzing a simple decision tree is straightforward and often can be done using popular software such as Microsoft Excel. Moreover, decision trees with added complexity can be developed and analyzed by more specialized software such as TreeAge (Shouman et al. 2011). However, decision trees can be restrictive when it comes to modeling non-linear relationships and long-term outcomes.

2.1.2 Markov Model

Markov models are a more advanced approach than decision tree models and have been used to tackle more complicated questions in health decision sciences (Yaylali et al. 2014). Markov models include Markov chains, which can be represented by two components: an initial state vector and a transition matrix. The transition matrix is a matrix whose entries are given by the probabilities of transitioning from one state to another. Multiplying the distribution of the previous state to the transition matrix gives the distribution of the current state. The states included in the Markov model must be mutually exclusive and collectively exhaustive. This class of models are rooted in the assumption that future states depend only on the current state (known as the Markov or "memoryless" property) (Oguzhan et al. 2010). This property simplifies the calculations significantly, reducing the computational burden of the simulation. A Monte Carlo method is often used along with Markov chain to capture parameter uncertainty from a

predetermined probability distribution (Geyer 1992). Markov models used in health decision sciences are usually discrete-time stationary Markov chains.

Unlike decision tree models, Markov models are more flexible in that they are conducive for modeling recurrent events and long time horizons. However, the "memoryless" property of Markov chains makes it difficult to capture individual health histories. For example, it is difficult to capture the increased mortality rate with the number of stroke an individual previously experienced using Markov models. Also, Markov models are cohort-based models that only capture population-level health outcomes and cannot study the effect of interventions across heterogeneous populations.

2.1.3 Microsimulation

Microsimulation model (aka. individual-level model or first-order Monte Carlo simulation) tracks individual-level health behaviors and outcomes (Rutter et al. 2011; Jalal et al. 2017; Krijkamp et al. 2018). Hence, it is easy to track an individual's health history and study heterogeneous populations. Also, in a microsimulation model, transition probabilities can depend on both population characteristics and the history of health states. For example, when modeling cardiovascular disease (CVD) with microsimulation, the incidence of future CVD can be dependent on both individual demographics (e.g., age, sex, race) and the history of cardiovascular health risk factors (e.g., high blood pressure, diabetes, cholesterol level). This gives microsimulation models greater flexibility in modeling details in the development of chronic disease compared to Markov models. A downside of microsimulation models is that they are more computationally expensive compared to Markov models. Also, microsimulation models are unable to capture interactions between individuals.

2.1.4 Agent-Based Modeling

Agent-based modeling (ABM) is a bottom-up modeling approach used to understand real-world systems when the behaviors and interactions of individuals and their environment are important (Railsback and Grimm 2012). Usually ABM is built on simple rules of behavior and action, which is then assessed against reality. ABM is popular among social science researchers because of its added realism through modeling interactions between individuals and with its environment.

ABM can be used to model social and health systems in an intuitive way that is appealing to policymakers. However, with the added complexity, ABM usually requires a more complicated structure and more comprehensive individual-level data compared to Markov models (Li et al. 2016). ABM also entails more computational resources, which often requires model developers to make a difficult trade-off between model complexity and computational efficiency.

2.2 Programming Languages and Software

The choice of programming language and software plays an important role in model development in health decision sciences. Based on June 2020 TIOBE programming community index, the following are the top 20 most popular programming languages from top to bottom: C, Java, Python, C++, C, Visual Basic, Java Script, PHP, R, SQL, Swift, Go, Ruby, Assembly language, MATLAB, Perl, PL/SQL, Scratch, Classic Visual Basic, and Rust (TIOBE). This index is based on ranking of how frequently the programming language shows up in search engine queries. While this index is an imperfect measure, it provides insights about changes in popularity of programming languages. However, the ranking is measured among modelers across all fields and is not specific to health decision sciences. It is possible that health decision scientists would have different preferences in programming languages.

Several programming languages are popular among health decision scientists, including Java, MATLAB, C++, Python, and R among others. In addition, some specialized software (e.g., TreeAge, AnyLogic) are also popular. Specialized software usually have user-friendly interface and are easier to learn, which make them ideal tools for health researchers who may have limited training in computer programming. In

comparison, programming languages such as C++, Python, and R provide more flexibility in developing decision science models and they are supported by large communities of professionals and researchers.

Many health decision scientists use R to develop models because of familiarity. Health researchers are often introduced to R through taking courses in epidemiology and biostatistics. The popularity of R is mainly attributed to it being free, having a huge online community, and having a great searchable archive such as on Stack Overflow. The CRAN website also makes many useful packages accessible. While R may not be the most computationally efficient programming language, it is easier to learn compared to other programming languages such as C and Java. This would allow modelers to start creating models sooner without committing extensive time to learn the complex syntax for a new programming language. Therefore, establishing a good practice to develop simulation models using R can help more researchers in health decision sciences delve into simulation modeling to add to their research objectives.

On a separate note, software such as TreeAge and Microsoft Excel are often used by health decision scientists to develop simulation models. Generally, it is easier to use TreeAge and Excel to develop models, but they are restricted by the functionalities developed into the platform. Likewise, there are a plethora of other software that were built with specific model structures in mind. These are great options when they suit the modeling needs of the researcher, but they greatly sacrifice flexibility. Moreover, there are some concerns about using underdeveloped software. For example, the RAND() function in Excel is a pseudo-random number generator that relies on a formula. While the pseudo-random number is not problematic for simple scenarios, its reliability may be of concern in more complicated models. Generally, using software developed for a specific type of modeling can make developing a model easier and faster but it can also be restrictive because of its lack of generalizability and transparency.

3 AN EXAMPLE OF MICROSIMULATION DEVELOPED USING R: THE SIMULATION FOR HEALTH IMPROVEMENT AND EQUITY (SHINE) MODEL

In this section, we describe a recent microsimulation model (the SHINE Model) that was developed using R. While we chose to refrain from getting into the details of how the simulation was coded on R, we hope to set an example of model development processes in R.

3.1 SHINE Model Background

The SHINE Model was developed by a multidisciplinary team consisting of simulation modelers, health economists, health policy researchers, and cardiovascular epidemiologists and physicians. The goal of the SHINE Model is to evaluate the impact of different public health policies in preventing cardiovascular disease (CVD). Approximately, one in every four deaths in the United States is due to CVD related reasons (Centers of Disease Control and Prevention) and racial/ethnic minorities are particularly vulnerable to CVD and CVD related mortality. Moreover, 80% and 50% of deaths due to CHD and stroke, respectively, are preventable with lifestyle or environmental adjustments (American Heart Association ; Cleveland Clinic). This motivates interests among health decision scientists to build models to inform policies to better prevent CVD and reduce health disparities.

The structure of the SHINE Model follows the well-established Cardiovascular Disease (CVD) Policy Microsimulation Model (Kohli-Lynch et al. 2019). The CVD Policy Microsimulation Model focuses on simulation of national policies and interventions that can be used to prevent and manage CVD. However, many public health policies are implemented at the local level. Hence, it motivated the effort to create the SHINE Model which simulates the impact of public health policies on people in NYC. Along with the shift to city-level analysis, the simulation model moved from TreeAge to R. There were a few reasons for such transition: increased flexibility in parameters and output structure, better transparency in computation, and reduced software cost. However, transitioning the model to R created a steeper learning curve for others who wanted to work with the model.

3.2 SHINE Model Structure

The SHINE Model is composed of health states and transition probabilities, as shown in Figure 1. Each health state is associated with an outcome measure (e.g., cost and QALY) that accumulated over time and are assessed against other scenarios. The transition probabilities are represented by the arrows in Figure 1. They represent the probability or risk of an individual moving from one state to another.

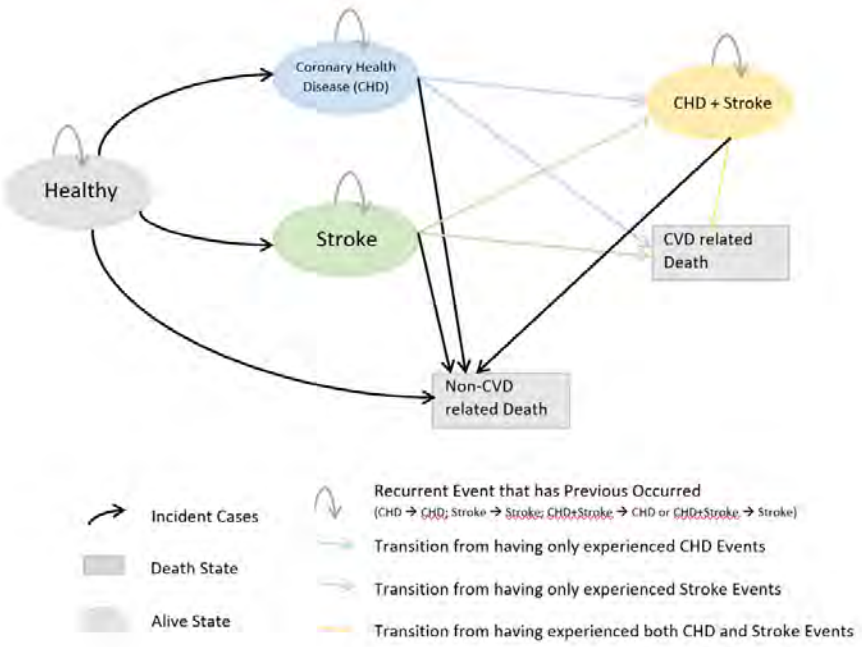

Figure 1: The SHINE Model Structure

The CVD-related events are categorized into two subcategories in the model: coronary heart disease (CHD) and stroke. CHD includes angina, coronary artery disease (ST-Elevation Myocardial Infarction (STEMI) and Non-ST-elevation myocardial infarction (NSTEMI)), and heart failure. Therefore, there are six main health states based on whether an individual has experienced: no CVD related events (Healthy), one or more CHD events (CHD), one or more stroke events (Stroke), both CHD and stroke at least once (CHD + Stroke), a Non-CVD related death, or a CVD related death. Each health state has a corresponding cost and quality-adjusted life year (QALY), which are accumulated as the simulation progresses.

The model can estimate lifetime survival rates, CVD related events, quality-adjusted life years (QALY), and medical costs associated with an intervention.

3.3 Simulated Population

Since a microsimulation model captures individual-level health outcomes, it is important to use individual-level data as model input. For the SHINE Model, individuals are sampled from the New York City Health and Nutrition Examination Survey (NYC HANES) (NYCHANES) to represent the NYC population. NYC HANES is similar to the National Health and Nutrition Examination Survey (NHANES) (Centers for Disease Control and Prevention) but only focuses on adults in NYC. It is a health survey conducted by the NYC Department of Health and Mental Hygiene to identify health factors and track improvements. The first NYC HANES data were collected in 2004 with a sample size of approximately 2,000 people.

The second NYC HANES data were collected in 2013-2014. Individuals were sampled according to the post-adjusted sample weights to be representative of NYC adults.

3.4 Transition Probabilities

The transition probabilities in the SHINE model were either calculated using data-driven incidence equations or estimated from the published literature. The probabilities out of the "Healthy" state and probability of non-CVD death were calculated using individual-level characteristics. We will refer to these probabilities as the incidence probability for simplicity. The equations to calculate these probabilities will be referred to as the incidence equations, shown as equation (1) below. All remaining probabilities were estimated from published literature.

Determined by the transition probabilities, simulated individuals move from one state to another as simulation runs. Figure 2 shows how the probabilities can be organized using R.

	P(Healthy)	P(CHD)	P(Stroke)	P(CHD + Stroke)	P(CVD-related Death)	P(non-CVD related Death)
1	0.5	0.2	0.14	0.05	0.03	0.08
2	0	0	0.5	0.25	0.15	0.1
3	0	0.6	0	0.2	0.1	0.1
4	0.75	0.08	0.05	0.005	0.005	0.11
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure 2: Example of a Transition Probability Layout

This transition probability matrix in Figure 2 feeds into a function that determines each individual's health state in each cycle. Krijkamp et al. (Krijkamp et al. 2018) provides pseudo code for such a function. In addition, the model updates the transition probabilities during each cycle depending on the following variables (when data is available):

- Current health state
- Race/ethnicity
- Length at each state
- Age (each cycle represents a year)
- Body mass index
- Smoking status (former smoker, current smoker, cigarettes per day)
- Systolic blood pressure
- Diabetes status
- HDL-c and LDL-c
- Estimated glomerular filtration rate (eGFR)

These variables are time-varying and future values are projected by matching the NYC HANES participants with the National Heart, Lung, and Blood Institute (NHLBI) Pooled Cohorts dataset. The approaches behind these variable projections are described in detailed in Hazzouri et al. (2019) and Zhang et al. (2019).

This variability is captured by the incidence equation mentioned previously. The incidence of a primary CVD event and the probability of non-CVD death is calculated using equation (1). The parameters (α and β's) are estimated using the Cox proportional hazards functions based on the NHLBI Pooled Cohorts dataset (Zhang et al. (2019)).

$$P(event) = \frac{exp(\alpha + \beta_{BMI}X_{BMI} + \beta_{LDL-c}X_{LDL-c} + ... + \beta_{sbp}X_{sbp})}{1 + exp(\alpha + \beta_{BMI}X_{BMI} + \beta_{LDL-c}X_{LDL-c} + ... + \beta_{sbp}X_{sbp})} \quad (1)$$

The Pooled Cohorts dataset includes data from Atherosclerosis Risk in Communities (ARIC), Coronary Artery Risk Develop in Young Adults (CARDIA), Cardiovascular Health Study (CHS), Framingham Heart Study – Offspring Cohort (FHS-O), Health, Aging, and Body Composition (HABC), Hispanic Community Health Study/Study of Latinos (HCHS/SOL), Jackson Heart Study (JHS), Multi-Ethnic Study of Atherosclerosis (MESA), and Strong Heart Study (SHS) (Oelsner et al. 2018; Zhang et al. 2019).

Finally, the SHINE Model can capture recurrent events. For example, if a simulated individual experiences an initial episode of CHD or stroke, that individual is at risk for a second (recurrent) CVD event or CVD-related death. We assumed that a maximum of two events is allowed during each cycle. It is worth noting that there is a nuanced distinction between being in a specific state and experiencing a specific event. For example, an individual who has previously experienced stroke is in the stroke health state but may not have had an episode of stroke in a long time. This distinction of a stroke event and stroke state (having experienced at least one episode of stroke previously), is important mainly because of differences in medical cost.

3.5 Model Assumptions and Other Considerations

There are a number of model assumptions in the development of microsimulations, which are summarized below.

- All the simulated individuals start from the initial state of "Healthy" as the model runs. We have removed individuals with any history of CVD.
- The model follows a group of individuals and therefore no one enters the simulation.
- Health care cost and QALY are dependent on both health state or the individual's age, gender, and length of state (wherever enough evidence is available).

As noted previously, the main purpose of a microsimulation model is to inform policies to improve population health and reduce health disparity. In the SHINE Model, the policy effect is usually captured by changing the incidence probability of transitioning to the two disease states (CHD and Stroke). All these details were identified as parameters that were important and supported by existing evidence from published literature. As it will be discussed in the next section, it is important to identify the details of the model for a streamlined model development process. Especially when the model involves a lot of details, it becomes challenging to develop, interpret, and collaborate on models without an organized structure.

Lastly, sensitivity analysis should be conducted once the SHINE Model is developed. Sensitivity analysis helps modelers and stakeholders understand how sensitive the results of the model are to specific model parameters. That being said, it would be impractical to conduct the sensitivity analysis against all the model parameters. A good practice for sensitivity analysis is to examine those parameters that are most uncertain or are most important to simulation results.

4 RECOMMENDATIONS ON THE DEVELOPMENT OF SIMULATION MODELS USING R

This section provides guidance and resources for developing microsimulation models using R. It is by no means comprehensive and is only meant to serve as a brief introduction for those learning to use R beyond its statistical capabilities.

4.1 Preparation: brainstorming and organizing

As with any coding languages, it is important for R modelers to design the model structure and specify input and output variables before writing any code. With a clear model structure, modelers can better visualize and structure the organization of the code. An organized code makes it easy to edit, debug, and share with collaborators. Modelers should consider creating the following items prior to coding:

- Model Diagram

- Underlying assumptions
- Input and output variable data structure and parameters
- Sketch of figures and tables for result presentation

Once the model structure and variables are fixed, workspace organization should be considered before any codes are written. Unlike codes written for the purpose of statistical analysis, the codes developed for simulation models are usually lengthier and have a more complex structure and organization. This includes figuring out how all the files will be stored relative to each other. Organizing your workspace also helps when sharing the model, especially when the model has external files as inputs. Generally, when working in an academic setting, having model inputs imported from an external file such as in the CSV file is preferred. We would recommend this, especially when working with people from different fields who may not be as comfortable with R. The organization of the code can be maintained through proper annotation. Dividing portions of the code with "sections" can also help create a clear structure in the code. If the code is extremely long, breaking up subsections into separate R files may also be beneficial. Separate R files can be read in using the function `source("file name")`.

4.2 Utilize existing R packages

One of the major strengths of R is that many packages are available to the public for free. These packages can significantly simplify the coding process. Packages are a collection of functions, which contain commands that accomplish a specific task given specific inputs. The R environment comes with a set of pre-installed packages (i.e., default packages). More packages are easily accessible from the CRAN website using the command `install.packages("Package Name")` followed by `library("Package Name")`. Additional packages can be also found from websites such as Bioconductor and Github through BiocManager and devtools packages, respectively.

We describe several packages that are specific to microsimulation and simulation modeling. For example, the `clinDR` package can be used for simulating and analyzing clinical dose response models. The `DES` and the `simmer` packages can be used for developing discrete-event simulations. The `ESGtoolkit` package can be used for conducting Monte Carlo simulations. The `MicSim` and `Sms` packages can be used for performing continuous-time microsimulation and spatial microsimulation, respectively. In addition, the `hesim` package can be used for conducting health-economic simulation modeling and decision analysis. Other more general-purpose packages can also be useful in developing simulation models. For example, `tidyverse` is a popular set of packages for any data science related applications (including simulation modeling).

4.3 Create customized functions using R

A modeler can develop his/her own function if a function/package is not readily available. Using the development of the SHINE Model as an example, we created a function called `Consumption()` to generate a sampled fruit and vegetable consumption level with ease. The function is useful because it pulls the necessary consumption data statistics for a given race and gender and use these data to estimate the parameter of the gamma distribution. We could then sample individual-level consumption from the gamma distribution. The following function can be called as follows:

```
Consumption(AA, Female, Fruit\_Consumption)
Consumption <- function(race, gender, DATA){
  # Initialize the output matrix
  pop_consumption <- matrix(NA, nrow = length(gender), ncol = 1)
    for (r in unique(race)){
      for (a in unique(gender)){
        # Input the statistics
        con_mean<- DATA$mean[DATA$race == r & DATA$gender == a]
```

```
        con_sd <- DATA$sd[DATA$race == r & DATA$gender == a]
        # Estimate the gamma distribution parameters
        shape_V = (con_mean)/con_sd # beta
        rate_V = (con_mean)^2/con_sd # alpha
        # Sample from the gamma distribution
        pop_consumption[(race == r & gender == a)] <-
            rgamma(n = length(pop_consumption[(race == r & gender == a)]),
                   shape = shape_V,scale = rate_V)}}
# Output the output matrix
return(pop_consumption)}
```

4.4 Understand data structure

It is important to understand different data structures commonly used in R before developing a microsimulation model. Some functions require specific data structure and would cause errors otherwise. The most common data structures in R are atomic vector, matrix, array, list, and data frame. Atomic vectors and lists are 1-dimensional, matrices and data frames are 2-dimensional, and arrays are 3-dimensional. Atomic vectors, matrices, and arrays are considered homogeneous while lists and data frames are considered heterogeneous. Homogeneous data structure means all the content in the data must be of the same type (e.g., numeric). Heterogeneous data structure instead allows the content to have different types (e.g., numeric, factors, characters). Working with the proper data structure avoids unnecessary complications as the model development moves forward.

For simulation model development, matrices, arrays, and data frames are the most useful data structures. It allows related data to be stored under one variable. Moreover, often the functions in the `tidyverse` package require the data to be in a data frame structure. Apart from `tidyverse`, in order to easily manipulate these data structures, a function called `apply()` is also valuable. The function takes a minimum of three inputs: `apply(X, MARGIN, FUN,...)`. The input `X` is the data and `FUN` is the function that should be applied to the data, such as sum. The `MARGIN` input is what makes this function versatile. For example, given `FUN=sum`, when `MARGIN=1` the function sums the entries along the rows, when `MARGIN=2` the function sums the entries along the columns, and when `MARGIN=c(1,2)` the data is added along row and column. This versatility comes in handy especially when summarizing results from multiple iterations. Results from different iterations are usually stored in arrays and calculating summary statistics can become inefficient without `apply()`.

Even without the `apply()` function, modelers should still try to vectorize their data whenever possible. For a simple example, take the task of adding two lists of numbers. The inefficient way is to add the two lists of numbers using a for loop. Instead, one should always use vector addition, which is significantly more efficient. This does not only apply to modeling in R, but any other programming languages, to improve code efficiency and reduce computational time.

5 PITFALLS AND FUTURE RESEARCH DIRECTIONS

Although developing microsimulation models using R is a promising direction in health decision sciences, it has several disadvantages and pitfalls that should be understood by modelers of interest. We describe these pitfalls and potential research directions to address them below.

First, although microsimulation can capture changes of health outcomes and behaviors at the individual level, it cannot capture interactions among individuals. It is thus difficult to use microsimulation to model infectious diseases in which disease transmission among individuals prevails or addictive behaviors such as smoking in which peer influence plays an important role. As described in Section 2, agent-based modeling would be more suitable to study interactions among individuals and how these interactions may affect population health outcomes.

Second, there are obstacles for modelers without extensive training in computer programming to develop microsimulation models using R. Although R provides numerous resources (e.g., packages, tutorials) that can ease the model development process, it is not a visualized model development tool. For health decision scientists who may have only received limited training in computer programming, the ability to develop simulation models in a visualized, user-friendly platform is important. Efforts should be taken by researchers to further ease the model development process and make microsimulation modeling with R more accessible to health decision scientists.

Third, it may be difficult to communicate a microsimulation model developed using R or other high-level programming languages with health care decision-makers or policymakers. A well-designed model diagram will be helpful to facilitate the communication between modelers and non-modelers and help policymakers understand the model structure. However, there are no guidelines for developing such diagrams and the quality of diagram varies across different model applications. Research is needed to standardize the construction of model diagrams and make it an essential component in the model development process.

Lastly, there needs to be a systematic way to determine which values to use as model parameters when multiple data sets or sources provide different estimates. The lack of such a systematic approach creates ambiguity in the choice of input values for the model and, thus, inevitably increases the uncertainty in the simulation result. A future direction is to link the meta-analysis capacity of R with the development of microsimulation models to provide a systematic process for parameter estimation and improve the robustness and reliability of the simulation results.

6 CONCLUSION

As more health care data become available and the need for better allocation of scarce health care resources grows, health decision sciences become critical in advising health care administrators and policymakers. Models cannot replace well collected data, but it can add to our understanding of those data sets and provide policy insights that traditional statistical models are unable to provide. Although microsimulation models and R may not be the most sophisticated tools in health decision sciences, they provide modelers with a convenient option to assess the impact of different policies and programs on patient outcomes and health care costs. Developing a microsimulation model may seem like a daunting task at the beginning. However, as described in the current paper, by following a structured procedure and taking advantage of existing resources, modelers may find the task more manageable. Research is needed to further strengthen the link between microsimulation models and R and make the modeling process more convenient even for researchers without a modeling background.

ACKNOWLEDGMENTS

Research reported in this publication was supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under Award Number R01HL141427. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

American Heart Association. "CDC Prevention Programs". https://www.heart.org/en/get-involved/advocate/federal-priorities/cdc-prevention-programs, accessed 05.18.2020.

Centers for Disease Control and Prevention. "National Health and Nutrition Examination Survey". https://www.cdc.gov/nchs/nhanes/index.htm, accessed 04.24.2020.

Centers of Disease Control and Prevention. "Heart Disease Facts". https://www.cdc.gov/heartdisease/facts.htm, accessed 04.20.2019.

Cleveland Clinic. "Stroke Risk Factors & Stroke Prevention". https://my.clevelandclinic.org/health/articles/13398-know-your-risk-factors-for-stroke, accessed 07.11.2020.

Geyer, C. J. 1992. "Practical Markov Chain Monte Carlo". *Statistical Science* 7(4):73–83.

Hazzouri, A. Z. A., E. Vittinghoff, Y. Zhang, M. J. Pletcher, A. E. Moran, K. Bibbins-Domingo, S. H. Golden, and K. Yaffe. 2019. "Use of a Pooled Cohort to Impute Cardiovascular Disease Risk Factors across the Adult Life Course". *International Journal of Epidemiology* 48(3):1004–1013.

Hunink, M. G. M., M. C. Weinstein, E. Wittenberg, M. F. Drummond, J. S. Pliskin, J. B. Wong, and P. P. Glasziou. 2014. *Decision Making in Health and Medicine: Integrating Evidence and Values*. 2nd ed. University Printing House, Cambridge, UK: Cambridge University Press.

Jalal, H., P. Pechlivanoglou, E. Krijkamp, F. Alarid-Escudero, E. Enns, and M. M. Hunink. 2017. "An Overview of R in Health Decision Sciences". *Medical Decision Making* 37(7):35–46.

Kamiński, B., M. Jakubczyk, and P. Szufel. 2018. "A framework for sensitivity analysis of decision trees". *Central European Journal of Operations Research* 26(1):135–159.

Kohli-Lynch, C. N., B. K. Bellows, G. Thanassoulis, Y. Zhang, M. J. Pletcher, E. Vittinghoff, M. J. Pencina, D. Kazi, A. D. Sniderman, and A. E. Moran. 2019. "Cost-Effectiveness of Low-Density Lipoprotein Cholesterol Level–Guided Statin Treatment in Patients With Borderline Cardiovascular Risk". *JAMA Cardiology* 4(10):969–977.

Krijkamp, E. M., F. Alarid-Escudero, E. A. Enns, H. J. Jalal, M. G. M. Hunink, and P. Pechlivanoglou. 2018. "Microsimulation Modeling for Health Decision Sciences Using R: A Tutorial". *Medical Decision Making* 38(3):400–422.

Li, Y., M. A. Lawley, D. S. Siscovick, D. Zhang, and J. A. Pagán. 2016. "Peer reviewed: agent-based modeling of chronic diseases: a narrative review and future research directions". *Preventing chronic disease* 13(69).

NYCHANES. "New York City Health and Nutrition Examination Survey". http://nychanes.org/about/, accessed 04.24.2020.

Oelsner, E. C., P. P. Balte, P. A. Cassano, D. Couper, P. L. Enright, A. R. Folsom, and J. Hankinson. 2018. "Harmonization of Respiratory Data From 9 US Population-Based Cohorts The NHLBI Pooled Cohorts Study". *American Journal of Epidemiology* 187(11):2265–2278.

Oguzhan, A., H. Hsu, A. J. Schaefer, and M. S. Roberts. 2010. "Markov Decision Processes: A Tool for Sequential Decision Making under Uncertainty". *Medical Decision Making* 30(4):474–483.

Railsback, S. F., and V. Grimm. 2012. *Agent-Based and Individual-Based Modeling: A Practical Introduction*. 2nd ed. 41 William Street, Princeton, NJ, USA: Princeton University Press.

Rutter, C. M., A. M. Zaslavsky, and E. J. Feuer. 2011. "Dynamic microsimulation models for health outcomes: a review". *Medical Decision Making* 31(1):10–18.

Shouman, M., T. Turner, and R. Stocker. 2011. "Using decision tree for diagnosing heart disease patients". *CRPIT* 121(9):23–30.

TIOBE. "TIOBE Index for July 2020". https://www.tiobe.com/tiobe-index/, accessed 07.20.2020.

Yaylali, E., J. S. Ivy, and J. Taheri. 2014. "Systems Engineering Methods for Enhancing the Value Stream in Public Health Preparedness: The Role of Markov Models, Simulation, and Optimization". *Public Health Reports* 129(4):145–153.

Zhang, Y., E. Vittinghoff, M. J. Pletcher, N. B. Allen, A. Z. A. Hazzouri, K. Yaffe, and P. P. Balte. 2019. "Associations of Blood Pressure and Cholesterol Levels During Young Adulthood With Later Cardiovascular Events". *Journal of the American College of Cardiology* 74(3):330–341.

AUTHOR BIOGRAPHIES

HEESUN EOM in a Data Analyst in the Department of Health Population Health Science and Policy at the Icahn School of Medicine at Mount Sinai. She holds a MS in Biomedical and Molecular Nutrition from Gerald J. and Dorothy R. Friedman School of Nutrition Science and Policy at Tufts University. Her BS is in Chemical Engineering and Statistics from Robert R. McCormick School of Engineering and Applied Science. She has been involved in assessing different health and nutrition policies from the perspective of cardiovascular disease and cancer outcomes. Her email address is heesun.eom@mountsinai.org.

YAN LI is an Associate Professor in the Departments of Population Health Science and Policy, and Obstetrics, Gynecology, and Reproductive Science at the Icahn School of Medicine at Mount Sinai. He is also the Director of the Health Policy Modeling Laboratory and a member of the Blavatnik Family Women's Health Research Institute at Mount Sinai. He is a Fellow of the New York Academy of Medicine. His research interests include systems science modeling in population health and health care management, cost-effectiveness analysis, and big data analytics. His email address is yan.li1@mountsinai.org.