Evolution of Simulation Modeling in Health Policy over a Half Century

Mohammad S. Jalali, MS, PhD\textsuperscript{1,2,*}, Catherine DiGennaro\textsuperscript{1}, Abby Guitar\textsuperscript{2}, Karen Lew\textsuperscript{2}, and Hazhir Rahmandad, PhD\textsuperscript{2}

\textsuperscript{1} MGH Institute for Technology Assessment, Harvard Medical School, Boston, MA, 02114
\textsuperscript{2} Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, 02139

Corresponding author:
Mohammad S. Jalali, MS, PhD
Assistant Professor, Harvard Medical School
Senior Scientist, Massachusetts General Hospital
msjalali@mgh.harvard.edu | (617) 724-3738

Word count:
2,967
Evolution of Simulation Modeling in Health Policy over a Half Century

Mohammad S. Jalali, MS, PhD1,2,*, Catherine DiGennaro1, Abby Guitar2, Karen Lew2, and Hazhir Rahmandad, PhD2

1 MGH Institute for Technology Assessment, Harvard Medical School, Boston, MA, 02114
2 Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, 02139

ABSTRACT

OBJECTIVE
To provide an overview of applications of simulation models in health policy, analyze the use of best reporting practices, and assess the reproducibility and quality of studies.

DESIGN
Systematic review of studies that used simulation modeling to inform health policy.

DATA SOURCES
Web of Science and PubMed until March 2016.

ELIGIBILITY CRITERIA FOR STUDY SELECTION
Studies that used simulation modeling as the core method to address any health policy questions.

MAIN OUTCOME MEASURES
Reproducibility was assessed using predefined, categorical criteria. Health policy domain distribution and changes in quality over time were well-characterized using MeSH terms and model characteristics, respectively.

RESULTS
This systematic search identified 1,613 eligible studies. A subset of 100 articles (50 top-cited and 50 randomly selected) were selected to analyze in-depth criteria for reproducibility and quality. We found an exponential growth in simulation-based studies of health policy over the past half century, with the highest growth in dynamic modeling approaches. The largest subset of studies is focused on disease policy models (70%), within which pathological conditions, viral diseases, neoplasms, and cardiovascular diseases account for about one-third of the articles. Nearly half of the 1,613 articles do not report the details of their models. Significant gaps against modeling best practices could be found in both random and highly-cited samples; only seven of 26 in-depth evaluation criteria were satisfied by more than 80% of samples. We found no evidence that the highly-cited sample of articles is better at following modeling best practices, defined as reporting model equations, specifying modeling assumptions, and discussing limitations, among others.

CONCLUSIONS
Simulation models are increasingly used to inform health policy, yet there is great variation in their adherence to best research practices. Our results suggest areas ripe for increased application of simulation modeling, as well as opportunities to enhance the rigor and documentation in the conduct and reporting of simulation studies in health policy.
Introduction
Increasingly complex health systems and advancing computational tools have promoted the application of simulation models to inform health policy. Simulation models help analyze and understand the complexity of a health issue, enabling the design, evaluation, and improvement of policies while minimizing unintended consequences [1].

Simulation applications span a wide range of areas such as health care reform [2], health care delivery [3], cancer research [4], and infectious diseases [5], among others. One objective of the current study is to provide a broad view of the state of simulation research across various health domains and modeling approaches to inform current trends and identify potential gaps.

A second objective of the current study is to provide a systematic assessment of the modeling practices in existing research. These models are integrated into operational thinking, policies, and interventions with significant impact. Hence, rigor and quality in design, implementation, validation, and dissemination are essential to the credibility and impact of simulation modeling research. However, enhancing quality partly relies on measuring quality attributes and monitoring progress, a task that has not yet been systematically pursued [6].

Our third objective is to assess the reproducibility of existing simulation research. Reproducibility is a cornerstone of science and calls for enhancing reproducibility have been on the rise [7-9]. Prior models should be revisited and revised as our understanding of real-world complexities evolves and new cases arise [10-12]. Documenting the state of the field informs the path forward and enables measurement of progress on this important issue.

We pursue these three objectives by conducting a systematic review of simulation research in health policy. The current review does not summarize relevant substantive findings; instead, our focus is on the trends in application areas, methods, and modeling practices.
Methods

Search strategy and selection criteria

As the primary search database, we used PubMed and searched for simulation and policy (and any variation of those key terms) in title and abstract. We complemented the results by searching 71 journals categorized in “Health Policy and Services” within Web of Science (WoS) for articles containing simulation in the title or abstract. We limited the search in both databases to English-language peer-reviewed articles. Our search included any articles indexed in these databases before March 2016. We reviewed the abstracts of the resulting sample to identify articles using simulation as the main method of research to answer a health policy question. Full text was inspected in cases where abstract did not establish this inclusion criteria. Studies which only mentioned simulation modeling and/or policy but did not employ simulation modeling as the main method were excluded in addition to reviews and meta-analyses. Figure 1 summarizes the search and inclusion/exclusion process.

Data extraction

We extracted the title, abstract, publication year, journal and author information directly from PubMed and WoS. To extract model type and more detailed reporting characteristics, we obtained and examined the full text of each article. To access the associated Medical Subject Headings (MeSH) terms from the articles selected for analysis, we developed a web scraping script in Python that was employed on PubMed in July 2018. We used a Python package (scholar.py) to scrape citation data from Google Scholar in October 2018.

Chronological, clustering, and trend analysis

We identified the categorizations of each MeSH term using the National Institute of Health’s MeSH Browser and determined the distribution of the full sample of articles across second-level categorizations (e.g., within Diseases [C], Neoplasms [C04]). In reporting the MeSH term frequencies below, we exclude the following categories because they were very general (and thus frequent) but not informative about policy areas of interest (we kept the articles containing them; each article may contain multiple second-level categorizations): Eukaryota, Amino Acids, Peptides and Proteins, and Hormones, Hormone Substitutes and Hormone Antagonists.

Using the same MeSH term frequency data, we present the distribution of all second-level MeSH terms across the articles. The four quadrants color coded to represent first-level categorizations within the MeSH terms include: Anatomy, Chemicals and Drugs, Diseases, and Organisms.

Finally, adopting from [13], we aggregated collaboration and location data for authors of each of the papers to determine a multilateral collaboration score (MCS) and related this metric with the results of modeling rigor and reporting evaluation. See the Supplementary Text for more information.

Categorization based on modeling approach

The review of titles and abstracts across all articles began with an evaluation of four high-level properties of models and their reporting: (1) static or dynamic (time-dependency), (2) stochastic or deterministic, (3) event-driven or continuous, and (4) model documentation, i.e., whether model equations are included in the paper/appendix, referenced in another paper, or not available. Two trained research assistants and the corresponding author conducted an initial
independent evaluation of 100 articles each and compared their results with each other in order to establish consistent evaluation criteria. The research assistants then completed the coding.

**Evaluation of modeling rigor and reporting**

After the broad overview and categorization of the full sample of articles, we selected 100 articles on which to perform an in-depth evaluation. These articles were selected by first identifying 50 highly-cited articles, and then picking 50 additional articles at random from those that remained. In identifying the 50 highly-cited articles, we controlled for the recency of the articles using a year fixed effect in a regression to predict expected publication; we then classified highly cited articles as those with largest fractional deviation from expected. The 50 highly-cited and 50 randomly selected articles are presented in Table S2 and Table S3, respectively.

Building on best modeling practice guidelines in the literature [14-16], we developed a set of 26 concrete criteria for evaluating the quality and rigor of simulation-based articles. These criteria spanned reporting of the model’s context, conceptualization, and formalization, along with analysis of the results and any external influences (see the criteria and their definitions in Table S1 and further detail on criteria assessment in the Supplementary Text). Coding criteria across two assistants were calibrated over an initial sample of 10 articles with the help of first author. Uncalibrated agreement rate on this initial sample across two coders was 75%. Once consistency in criteria was established the two completed the coding showing a 90% agreement level. The remaining 10% of split coded items were discussed and resolved; reported findings are based on the resulting consensus.

**Results**

**Study Identification and Selection (inclusion and exclusion criteria)**

Figure 1 presents the summary of study selection. We identified 5,092 articles from an initial search in the PubMed and WoS databases, of which 300 were duplicates found in both. We then reviewed the abstracts and titles of the resulting 4,792 articles. Of these, 1,855 articles met the initial criteria for full-text review. Next 272 papers were excluded following the full-text inspection, leaving 1,613 studies that contained simulation modeling as a core method in addition to a policy analysis (Figure 1). This is the full sample for the study.
Figure 1: Study Selection Flow

Chronological, clustering, and trend analysis

Figure 2 presents the breakdown of research areas. The broad majority of all research areas across articles fall under Diseases (37·5%), followed by Chemicals and Drugs (33·96%), Organisms (27·55%), and Anatomy (1·25%) (Figure 2). Each category is broken down into the most common subcategories. In Diseases, the most commonly occurring categories are Pathological Conditions, Signs and Symptoms, and Virus Diseases; in Chemicals and Drugs, Organic Chemicals is the most common; in Organisms, Viruses constitutes the vast majority of terms; in Anatomy, Body Regions (11 articles) and Cells (9 articles) are most common. See Figure S1 for the number of articles in the top 20 research areas. It should be noted that 5% (77 out of 1613) of the articles evaluated did not contain MeSH terms and were not included in Figure 2. It is important to note one article has the potential to contain MeSH terms more than one segment of the figure.

Categorization based on modeling approach

Figure 3 presents the trend of articles over time based on their modeling approaches. From 1967 to 2016, articles containing models that are both Static and Deterministic have seen the lowest increase in publication frequency, as shown by the exponential growth curve, followed by Static and Stochastic models. Dynamic and Deterministic models have seen a steady increase in publication, mostly accounted for between 2002 and 2016, while Dynamic and Stochastic models have the highest increase beginning between 1997 and 2001.
Figure 3: Number of published journal articles from each model type permutation between 1967 and 2016. Each article was categorized by type of model and sorted by publication date in five-year increments. The exponential growth rates are 15.6 for Static & Deterministic, 24.6 for Static & Stochastic, 48.5 for Dynamic & Deterministic, and 117.2 for Dynamic & Stochastic.

Figure 4 presents the trend in reporting of models from 1967 to 2016. Until 2007, the proportion of articles that did not report a model at all was greater than the ones that did. Between 2007 and 2016, more studies reported model equations than those that did not report. The overall percentage of studies that cite an earlier study with their model equations (9%) increases from 1967 to 2016, with the largest increase seen between 2007 and 2011.

Figure 4 also presents a breakdown of model reporting across the four model type permutations. 79% of articles containing models that are both Static and Deterministic (n=135) did not report their model equations, while 15% did within the text of the paper/appendix and 6% cited another article referencing the model equations. 44% of articles containing models that are both Static and Stochastic (n=193) did not report their model equations, while 52% did within the text of the paper/appendix and 4% cited another article referencing the model equations. 62% of articles containing models that are both Dynamic and Deterministic (n=370) did not report their model equations, while 30% did within the text of the paper/appendix and 8% cited another article referencing the model equations. 34% of articles containing models that are both Dynamic and Stochastic (n=868) did not report their model equations, while 55% did within the text of the paper/appendix and 11% cited another article referencing the model equations. Approximately, half of the articles were event-driven, and half were continuous, as shown in Figure S2.
Figure 4: Model reporting trend from 1967 to 2016 and in each permutation of model categorization. (a) articles were divided into five-year publishing date increments and then assessed by whether they contained model equations or citations of model equations. (b) each article was categorized by type of model and sorted by whether the model equation used was reported in the paper/appendix, in another paper, or not at all. The models were divided into the following four descriptive categories: static and (i) deterministic or (ii) stochastic, or dynamic and (iii) deterministic or (iv) stochastic. Size of pie charts indicates sample size of each category.

Three logistic regressions were performed to analyze whether articles reported their models (either directly or by citation to original model) based on whether it was static or dynamic, event-driven or continuous, and stochastic or deterministic, and whether model reporting was significantly impacted by the age of the article. Table S4 summarizes these results. Overall articles using static models are less likely to report those models than papers using dynamic models. In addition, event-driven and stochastic models are both more likely to be reported.

Evaluation of modeling rigor and reporting

Figure 5 presents whether the 100 sampled articles satisfy the 26 evaluation criteria. Of the 26 criteria, four had significantly higher rates of satisfaction in the random selection of articles: evaluation and testing ($p = 0.13$), generalizability discussion ($p = 0.09$), high-level model visualization ($p = 0.07$), and discussion about strategies/policies ($p < 0.001$). Two had significantly higher rates of satisfaction in the most-cited articles: quality of calibration fit ($p = 0.08$) and model calibration ($p = 0.01$) (Figure 5). Modeling code availability and reproducibility discussion were the lowest in both groups with only 1-2% of the articles from either group satisfying these criteria. Conversely, limitations discussion, assumptions, scope, objective, problem definition, parameter values and data sources, and the reporting of quantitative results were satisfied by the vast majority (>80%) of selected studies in both categories.
A principle component analysis suggests that the various criteria are not highly correlated and thus offer distinct measures for assessing the papers (See Table S5 for details). However, we found no strong predictor of aggregate quality measures. A linear regression was performed to analyze the overall evaluation score (maximum score=26 points) of the 100 sampled articles based on whether it was static or dynamic, event-driven or continuous, stochastic or deterministic, the journal impact factor, the number of citations, MCS, number of pages, time since publication, and the location of the authors’ affiliation (U.S. or non-U.S.). The article’s score increased only 0.07 points for each additional page (p=0.08). The other independent variables, including collaboration, number of citations received, and journal impact factor had no significant impact on the score (p>0.1). See Table S6 for the details of the regression analysis.

**Discussion**

In this study, we conducted a broad review of published journal articles that used simulation modeling to inform health policy to identify trends, document application domains, and assess modeling rigor and reproducibility. Between 1967 and 2016, this line of research has become more common and the type of models have shifted to better represent the complexity of real-world health issues, particularly within a dynamic and stochastic framework.

**Figure 5: Percentage of articles satisfying 26 in-depth evaluation criteria in four areas.** Each criterion was assessed for the 50 most-cited articles (square) and 50 random articles (circle). The percentage of articles from each group meeting the criteria are presented. Criteria marked with an * are significantly different between random and most-cited articles (p<0.1).
Our analysis also highlights the significant room for improvement in terms of reproducibility and quality of reporting of studies. We show that nearly half of studies do not report their model equations and that the most-cited articles perform no better than random articles when assessed against a host of evaluation criteria. The only (marginally) significant predictor of reproducibility was the length of the article (longer articles tended to perform better). Nevertheless, given the wealth of online repositories for storing models, codes, and data, page limit should no longer be a consideration in documenting models. On the other hand, factors such as collaboration, number of citations, and journal impact factor showed no significant impact on reproducibility. This goes against some common intuitions about the quality of research [17] and indicates a gap between perceived quality and actual reproducibility.

Our review also shows that dynamic models tend to be better reported than static models, and the same is true for stochastic and event-driven vs. deterministic and continuous models. Beginning in 2002, publications have moved toward creating dynamic and stochastic models to better reflect the real world. This time period also witnessed improvements in reporting model equations. We hope this trend will continue to accelerate as simulation models become more frequently relied upon in health policy decisions. We presented these gaps in model reporting and rigor in hopes that future publications meet a higher standard of developing and reporting simulation models and facilitate their reproducibility.

In our sample of simulation articles, the fraction of disease-focused articles is the largest. Areas such as pathological conditions and viral diseases receive a great deal of modeling attention, but parasitic diseases, for example, receive less modeling attention; although they affect hundreds of millions of people each year and cases are concentrated in tropical regions where populations have fewer resources [18]. Simulation modeling is well-suited to aid in understanding of complex issues where resources are limited. Within the context of global health, a gap in the literature is detrimental when considering potential interventions to prevent the spread of parasitic diseases, especially given that the projected global rise in temperature promotes the spread of these diseases [19-21]. There are other research areas with similar potential for applications of simulation models, for example chemically-induced disorders (under which opioid use disorder is studied) [22-24] and congenital and hereditary diseases, impact the population in increasingly complex ways [25] but are not prominent in current applications.

An important limitation of current study is our search strategy that focused on articles using the term “simulation” (and all its variants) in the title or abstract. We expect a sizeable number of articles exist that use simulation but do not use the term or its variants in the title and abstract, instead referring to various modeling approaches. Expanding the search strategy to include mathematical, computational, economic, or other modeling approaches would have enlarged the initial sample to tens of thousands of records, which would have been infeasible to review and analyze in depth. Considering this, we chose to review a large subset of the relevant literature that is explicit about using simulation, while acknowledging the bias inherent to excluding non-explicit simulation approaches.

This review may also be limited by selecting only peer-reviewed studies published in English and indexed on PubMed and WoS. We did not assess documents other than journal articles (e.g., conference articles or whitepapers), and a more comprehensive search strategy could include these sources. In addition, models in health policy may be developed privately, discussed in

11
institutional reports, or remain unpublished. Thus, the journal articles may be missing some of
the relevant results in simulation modeling.

For an in-depth analysis of the 100 articles, we devised a scoring system assigning one point to
each binary criteria met. This assignment only focuses on the existence of a criteria and does not
offer a more nuanced understanding of quality. We also do not include differential weights for
these criteria in assessing the overall quality measure. We acknowledge that those criteria vary in
their importance to overall quality and impact of research, yet, absent an objective method for
aggregating them, we preferred to avoid imposing our own subjective weights on various
criteria.

Further reviews can focus on ways to systematically analyze and improve the reproducibility of
simulation modeling. For example, they could investigate reproducibility across application
domains and levels of analysis (e.g., cell, individual, or society) to inform concrete suggestions
for specific communities of research. Additionally, future studies can compare our results on the
reproducibility of simulation models in health policy with other application domains of
simulation modeling. Tracking the same metrics over time provide another measure of progress.
Finally, using our dataset, machine learning methods may be trained to identify reproducibility
and quality metrics more efficiently and apply those criteria over larger bodies of research.

**Conclusions**

Our analysis highlights the changes in simulation study quality over the last half-century and
suggests several areas for improvement. Regardless of the quality of underlying model, lack of
reproducibility is a major flaw that erodes confidence in the policies and decisions these papers
inform and their broader impact, and thus deserves more attention in simulation research. We
hope that this study facilitates conversation around research gaps that can benefit from
simulation modeling and motivates modelers to collaborate in addressing those gaps and
increasing the diversity of research areas while enhancing the rigor and reproducibility of
simulation research in health policy.
Acknowledgments: We would like to express our gratitude to Professors Donald Burke (University of Pittsburgh), John Sterman (MIT), and Gary King (Harvard) who shared their suggestions and thoughts. We thank participants at Medical Decision Making and System Dynamics Conferences who provided constructive feedback on initial versions of this report. We also thank Meera Gregerson, Yikang Qi, and Yuan Yuan who assisted in data extraction and analysis at MIT.

Contributors: MSJ designed and conceived the project. AG and KL searched and reviewed the literature. CD and MSJ analyzed the data extracted from the reviewed articles. HR advised the review process. MSJ and CD wrote the first draft of the manuscript, and HR contributed to further development of the analysis and content. All authors reviewed the draft of the manuscript and provided comments and critical review.

Funding: No funding source was used to conduct this study.

Competing interests: The authors declare that there is no conflict of interest.

Ethical approval: Not needed.

Data sharing: The online appendix provides the details of the analysis. Additional information can be provided upon request from the corresponding author at msjalali@mgh.harvard.edu.

Transparency: The corresponding author affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

References:


