Volume 12 • Number 4 • 2009
VALUE IN HEALTH

Calibration of Disease Simulation Model Using an Engineering Approach

Chung Yin Kong, PhD,[1,2] Pamela M. McMahon, PhD,[1,2] G. Scott Gazelle, MD, MPH, PhD[1,2,3]

[1]Massachusetts General Hospital, Institute for Technology Assessment, Boston, MA, USA; [2]Harvard Medical School, Boston, MA, USA; [3]Harvard School of Public Health, Boston, MA, USA

ABSTRACT

Objectives: Calibrating a disease simulation model's outputs to existing clinical data is vital to generate confidence in the model's predictive ability. Calibration involves two challenges: 1) defining a total goodness-of-fit (*GOF*) score for multiple targets if simultaneous fitting is required, and 2) searching for the optimal parameter set that minimizes the total *GOF* score (i.e., yields the best fit). To address these two prominent challenges, we have applied an engineering approach to calibrate a microsimulation model, the Lung Cancer Policy Model (LCPM).

Methods: First, 11 targets derived from clinical and epidemiologic data were combined into a total *GOF* score by a weighted-sum approach, accounting for the user-defined relative importance of the calibration targets. Second, two automated parameter search algorithms, simulated annealing (SA) and genetic algorithm (GA), were independently applied to a simultaneous search of 28 natural history parameters to minimize the total *GOF* score. Algorithm performance metrics were defined for speed and model fit.

Results: Both search algorithms obtained total *GOF* scores below 95 within 1000 search iterations. Our results show that SA outperformed GA in locating a lower *GOF*. After calibrating our LCPM, the predicted natural history of lung cancer was consistent with other mathematical models of lung cancer development.

Conclusion: An engineering-based calibration method was able to simultaneously fit LCPM output to multiple calibration targets, with the benefits of fast computational speed and reduced the need for human input and its potential bias.

Keywords: algorithms, calibration, lung cancer, simulation model.

Introduction

Simulation modeling and clinical trials offer different and complementary methods of exploring the relationship between a health-care intervention and health outcomes. Clinical trials compare two or more clinical interventions and are able to describe the comparative effectiveness of the different options. Nevertheless, most clinical trials have a relatively short time horizon (2–7 years); by extrapolating from the short-term trial data, modeling can estimate the longer-term consequences (both positive and negative) of the intervention. In addition, clinical trials are very expensive and so rarely evaluate all the clinically relevant options; modeling can combine the results of several clinical trials, adjust for trial population differences, and estimate the incremental differences between options not compared in any one trial.

In some instances, simulation models incorporate unobservable natural history parameters to model disease development in patients. Often, estimates of logical relationships between observable parameters can be established using meta-analysis or evidence synthesis, but the values of unobservable natural history parameters must be obtained through the process of model calibration. Calibration is a process of varying the unobservable parameters until model outputs closely match existing clinical and epidemiologic data [1]. After the relevant epidemiologic data have been selected as calibration targets, calibrating the model to those targets consists of two parts: 1) defining how to simultaneously measure the level of discrepancy between model output and multiple calibration targets, and 2) searching for the parameter set which results in an overall minimization of that discrepancy.

Address correspondence to: Chung Yin Kong, Institute for Technology Assessment, 10th floor, 101 Merrimac Street, Boston, MA 02114, USA. E-mail: joey@mgh-ita.org

10.1111/j.1524-4733.2008.00484.x

Despite the importance of model calibration, there is no standard practice in the disease modeling literature for measuring the level of discrepancy between the model output and the calibration targets. Nor are there standards for how to search the parameter space such that the best parameter set is identified.

Methods of defining the level of discrepancy vary widely. Some researchers visually match the model output to the clinical data. This methodology is particularly problematic because it cannot be independently replicated: a different researcher using the same model may have selected a different parameter set.

Methods of searching the parameter space also vary widely. Some researchers manually vary input parameters in their models, while others use simple parameter search algorithms such as grid search [2,3] and random search [4]. Grid search divides each parameter value between the maximum and minimum values into regular grids. The results of all possible parameter set combinations are then compared to the calibration targets. The random search algorithm randomly picks input parameter sets within the allowable parameter space. Based on initial results, the allowable parameter region can be modified or narrowed, and searched again more thoroughly. Both methods attempt to locate the optimal parameter values by exhaustively exploring the whole parameter space. These methods will find the optimal parameter set but are only practical for simulation models with only a few parameters.

For a disease microsimulation model with many natural history parameters, grid and random search methods cannot sample the parameter space efficiently. Using a model with 20 unobservable parameters as an example, testing only 10 values of each unknown parameter with grid or random searches will require 10^{20} parameter sets. In practice, researchers would not test all 10^{20} sets; typically, results from searching one area of parameter space are visually inspected and interpreted before

selecting another area of parameter space to search. Nevertheless, even searching a portion of those possibilities would result in a time-consuming effort.

We sought to adapt calibration methodologies from the engineering literature that would enable an automated, computationally feasible, and time-efficient parameter search for comprehensive microsimulation models. Two fast parameter search algorithms commonly used in engineering simulation were identified as potentially meeting all of these criteria: simulated annealing (SA) [5,6] and genetic algorithm (GA) [7,8]. Both parameter search methods were applied to the Lung Cancer Policy Model (LCPM) [9–11], a microsimulation model of lung cancer development, progression, detection, treatment, and survival. In this article, we have provided examples of using these two fundamentally different optimization algorithms to perform the parameter search and compared their performance using relevant constraints and scenarios comparable to the process of initially calibrating a model, repeating the calibration of an individual component, or refining the calibration.

Material and Methods

LCPM

The LCPM is, as a comprehensive model of lung cancer, designed to evaluate screening [9–11] and other lung cancer control interventions. A detailed description of all components of the model is available online at the National Cancer Institute's (NCI) Cancer Intervention and Surveillance Modeling Network (CISNET) Web site (http://cisnet.cancer.gov/profiles/), but details relevant to the current study are provided here. The LCPM simulates cohorts of individuals with sex-, race-, and birth cohort-specific smoking histories as observed for the US population [12,13]. In each monthly cycle, a new cancer may develop in an individual, or an existing cancer may grow, or symptoms may develop. Lung cancers and benign pulmonary nodules can be detected through incidental imaging or scheduled screening. Cancers can also be diagnosed by an evaluation of a patient's symptoms. Patients with suspected lung cancer receive diagnostic and staging tests, and may receive surgical or nonsurgical treatment. Existing cancers may or may not be detected before an individual dies of their lung cancer or from another cause [14]. Smoking exposure is updated monthly. Model outputs include estimates of age-specific lung cancer incidence rates and distributions of lung cancer cell types and stage at diagnosis, as well as survival by stage at diagnosis.

To reflect known heterogeneity of lung cancer and allow the evaluation of a variety of interventions, the LCPM has a "deep" underlying natural history model, with cell type-specific parameters for growth, invasiveness, and relationship with smoking (among others). Simulating the development and progression of four cell types of lung cancer with such detail required calibrating the model to estimate values of 87 unobservable natural history parameters. In this study, we applied the optimization algorithms to a subset of 28 parameters that describe the development of new lung cancers (i.e., coefficients in the logistic equations for each cell type). The remaining 59 parameters related to the growth, progression, and symptom detection of existing cancers were not varied and their values were estimated in the previous calibration exercises.

Three cancers of any of four lung cancer cell types may develop in each simulated person (adenocarcinoma with or without bronchioloalveolar carcinoma, large cell, squamous cell, and small cell), which comprise over 90% of lung cancer. The monthly probability of developing the first malignant cell is calculated using an independent logistic equation for each cancer type. Each logistic function has a type-specific intercept, type-specific coefficients for *age*, age^2, years of cigarette exposure (smoke-years, *SY*), an interaction term between *SY* and age^2, the average number of cigarettes smoked per day (cigarettes per day, *CPD*), and the years since quitting (*YSQ*) smoking.

Before beginning the calibration procedure, we eliminated implausible parameter ranges and assumed correlations between some parameters on the basis of known relationships between the incidence of lung cancer by cell type and smoking history. For example, the baseline risks dictated by the type-specific intercepts were ordered to reflect the distribution of cell types among nonsmokers [15–18]. Lung cancer risk is also known to increase with age and smoking experience (*SY*), and is known to decrease as the *YSQ* increases [19–22]. The risk of small cell cancer is the most dramatically affected by smoking experience, and the effect of smoking cessation has the weakest effects for developing adenocarcinoma [17,23]. Thus, the type-specific coefficient for *SY* is required to have the largest magnitude for small cell cancer and the type-specific coefficient for *YSQ* is also required to have the smallest magnitude for adenocarcinoma.

To account for the changes in unmeasured risk factors (in addition to the change in smoking pattern) experienced by different birth cohorts, we incorporated one sex-specific birth cohort coefficient, β_{BY}, into the monthly probability of lung cancer development.

Calibration Targets

With the publication of new biological and medical findings, new calibration targets are constantly incorporated into our LCPM. The version of the LCPM used in this study had 11 calibration targets (nine primary and two secondary targets), derived from various data sources. The primary calibration targets were cancer incidence by cell type, stage-specific survival, and stage distribution at diagnosis. Primary targets were extracted from data from the NCI's Surveillance, Epidemiology, and End Results (SEER) Program [24]. The secondary calibration targets were age-specific mortality rates of nonsmokers and lung cancer-specific mortality ratios for current (vs. never) smokers, derived from past cohort studies [25] and other literature sources describing clinical experience [16,18,26,27]. All calibration targets in this study were derived from only publicly available de-identified human subject data.

The discrepancy between the simulation model output and each calibration target was measured by the goodness-of-fit (*GOF*) statistic. We calculated the measure of discrepancy between each LCPM output to the corresponding calibration target (*i*), GOF_i, using a sum-of-squared error *GOF* statistic (analogous to a chi-square statistic).

Method of Handling Multiple Targets

Calibrating the LCPM to multiple targets simultaneously required a definition of a global *GOF* statistic. We used the weighted-sum approach [28,29] in which weighting factors were assigned to all targets in advance of the calibration procedure to reflect the relative importance of the targets. Thus, the summary GOF_{sum} statistic is a linear combination of the individual statistics,

$$GOF_{sum} = \sum_i W_i \times GOF_i \quad (1)$$

where W_i is the weighting factor for the i^{th} target. For the LCPM, primary calibration targets were given a weight of 1.0 and secondary calibration targets were given a weight of 0.5 to avoid