ACRM
AMERICAN CONGRESS OF REHABILITATION MEDICINE

Archives of Physical Medicine and Rehabilitation

journal homepage: www.archives-pmr.org

Archives of Physical Medicine and Rehabilitation 2015;96:746-9

BRIEF REPORT

A Validation Study for Using Iterative Proportional Fitting to Weight the Traumatic Brain Injury Model Systems National Database: An NIDRR-Sponsored Study

Christopher R. Pretz, PhD,[a,b] Jeffrey P. Cuthbert, PhD,[a] Gale G. Whiteneck, PhD[a]

From the [a]Craig Hospital, Englewood, CO; and [b]Traumatic Brain Injury National Data and Statistical Center, Englewood, CO.

Abstract

Objective: To verify that iterative proportional fitting (IPF), or raking, has the desired effect of aligning estimates and parameters so that researches have confidence in population projections when weighting the Traumatic Brian Injury Model Systems National Database.

Design: Secondary data analysis using IPF.

Setting: Inpatient rehabilitation.

Participants: People aged 16 years and older with a primary diagnosis of traumatic brain injury receiving initial inpatient rehabilitation.

Intervention: Not applicable.

Main Outcome Measures: Age at injury, race, sex, marital status, rehabilitation length of stay, payer source, and motor and cognitive FIM scores.

Results: This study demonstrates the utility of applying IPF to weight the TBI Model System National Database so that results of ensuing statistical analyses better reflect those in the United States who are 16 years and older with a primary diagnosis of TBI and are receiving inpatient rehabilitation.

Conclusions: In general, IPF aligns population estimates on the basis of weighted Traumatic Brian Injury Model Systems data and known population parameters. It is reasonable to assume that IPF has the same effect on unknown variables. This provides confidence to researchers wishing to use IPF for making population projections in analyses.

Archives of Physical Medicine and Rehabilitation 2015;96:746-9

Iterative proportional fitting (IPF), also known as raking, was first proposed by Deming and Stephan[1] in the 1940s as a method of estimating individual cell probabilities in a contingency table using fixed marginal row and column totals as constraints. Over the years, this method has been adapted to generate weights that are used to calculate population estimates of unobserved variables on the basis of information from observed parameters, that is, known population proportion(s). IPF methods are used to weight numerous national survey studies, most notably the Centers for Disease Control and Prevention's Behavioral Risk Factor Surveillance Survey and the National Health Interview Survey. Similarly, the Traumatic Brain Injury Model Systems (TBIMS) National Data and Statistical Center, in coordination with the Centers for Disease Control and Prevention, has adopted IPF so that weights generated on the basis of known parameters can be used in analyses with the expectation that results are more representative of *all* late teens and adults who receive inpatient rehabilitation for a primary diagnosis of traumatic brain injury (TBI) in the United States.

Currently, the National Data and Statistical Center and the Centers for Disease Control and Prevention have used IPF in 4 TBI-related studies, with more planned and in progress.[2-5] Parameters used in these studies were obtained from aggregate data

Supported by the Traumatic Brain Injury (TBI) Model Systems National Data and Statistical Center, the National Institute on Disability and Rehabilitation Research (NIDRR) (grant no. H133A110006). The TBI Model Systems National Database is supported by the NIDRR and created and maintained by the TBI Model Systems Centers Program. This work was prepared at the Traumatic Brain Injury Model Systems National Data and Statistical Center, Englewood, CO. This article is intended to promote the exchange of ideas among researchers and policymakers. The views expressed in it are part of ongoing research and analysis and do not necessarily reflect the position of the U.S. Department of Education.

Disclosures: none.

0003-9993/14/$36 - see front matter © 2015 by the American Congress of Rehabilitation Medicine
http://dx.doi.org/10.1016/j.apmr.2014.09.025

on all late teens and adults admitted to inpatient rehabilitation for a primary diagnosis of TBI, which were submitted to 2 central data repositories that serve as intermediaries for the Centers for Medicare and Medicaid Services.[6] The 2 central repositories are the Uniform Data System for Medical Rehabilitation[7] and the American Medical Rehabilitation Providers Association database, eRehabData,[8] and both require submission of data from cases with both Medicare and non-Medicare payers. These data systems were approximated to include no less than 92% of all civilian rehabilitation facilities in the United States.[9] Because these rehabilitation facilities include the largest in the United States, this sample was considered to include close to 100% of all patients aged 16 years and older with a primary diagnosis of TBI receiving inpatient rehabilitation. The U.S. parameters (ie, the categorical distributions of specific variables) derived from these aggregate population data were age at injury, race, sex, marital status, rehabilitation length of stay, payer source, and motor and cognitive FIM scores. Missing data for these parameters ranged from 0.0% to 2.5% for the Uniform Data System for Medical Rehabilitation/eRehabData sample and 0.0% to 2.1% for the TBIMS data. The time period for this analysis included October 1, 2001, to December 31, 2007.

The purpose of this brief report was not to discuss the intricacies of IPF itself; for a more involved understanding of IPF, see the *Encyclopedia of Biostatistics* (2nd edition)[10] or the study by Kalton and Flores-Cervantes,[11] who also provide compelling evidence for using IPF. Instead, the purpose of this brief report is to supply information about the utility of IPF in producing population projections for variables within the TBIMS national database. Such information is timely and of significant importance as the interest in weighting the TBIMS data, and other national databases, becomes more prevalent.

Goals of this brief report were twofold. The primary goal was to determine whether the population parameters, that is, information that originates from the Uniform Data System for Medical Rehabilitation/eRehab data used in the IPF process, serve their purpose. That is, does use of the parameters "move" estimates of unknown variable(s) so that estimates and parameters become more aligned? A secondary goal was to provide a sense of which single variables are most influential in the IPF process and to compare results based on single variables with those based on multiple variables.

Methods

All analyses were performed using SAS (version 9.4).[a] Of course, the problem with achieving the primary goal was that one needs to know the population characteristics of a variable to observe the degree to which the value of the estimate and the parameter coincide. This problem is overcome by removing a single parameter from the IPF process and treating it as if it were unknown. In this process, remaining parameters are used to generate weights, which are, in turn, used to produce an estimate of the removed variable. Once the estimate is calculated, it is directly compared with the known parameter. This procedure is repeated, removing each variable in turn from the analysis, estimating, and subsequently comparing it with the known parameter.

List of abbreviations:

IPF	**iterative proportional fitting**
TBI	**traumatic brain injury**
TBIMS	**Traumatic Brain Injury Model Systems**

To determine which individual variables are most influential, we conducted a series of separate analyses, one for each variable, to provide estimates of the remaining variables. Estimates were then compared with their respective parameters by calculating the average and SD of the absolute value of the difference across the remaining variables, that is, the distance, where smaller values reflect an enhanced overall ability of a given variable to estimate remaining variables. Finally, the "overall" average distance, which was calculated by using the distance between estimates and parameters provided in table 1, was compared with the average distances generated from using individual variables.

Results

Results pertaining to the primary goal are provided in table 1 and indicate that in general, IPF functions as intended. That is, in most of the cases, IPF moves the TBIMS percentage toward its corresponding population parameter. The few cases in which the distance between the population estimate and the population parameter is larger than the distance between the TBIMS percentage and the population parameter are indicated with an asterisk.

Table 2 addresses the secondary goal and reports the average and SD of the difference between estimates and parameters for each of the remaining variables when the data were weighted on that single variable. Weights generated by age and payer source tend to produce estimates more in line with their respective parameters. However, even though age and payer source appear to be the best individual performers, on examining the distance between estimates and parameters provided in table 1, we find the average of these distances (the "overall" average distance) to be 2.95 ± 2.7, which highlights the benefit of using multiple variables in the IPF process. Table 2 also indicates that variables other than age and payer source produce similar means and SDs, though comparatively, these variables display a reduced capacity in providing good estimates when used individually.

Discussion

In analyzing the TBIMS national database, IPF demonstrated utility in aligning estimates and parameters. The known population parameters were more closely estimated by the IPF process than by the TBIMS national database in 33 of 38 categories across 8 variables. Although the degree to which the alignment is deemed sufficient may be debated, it is evident if population projection is the intent; using IPF is more advantageous than not using IPF. Such knowledge should bring comfort to researchers whose desire is to weight the TBIMS national database when conducting analyses in which the intent is to project characteristics and outcomes to the larger population of those 16 years and older who received inpatient rehabilitation with a primary diagnosis of TBI.

The finding that both age and payer source were the most influential individual factors in projecting to the larger population may be due to the confounding of payer source with age (ie, the vast majority of people covered by Medicare are older than 65y) and is consistent with the previous finding that the distribution of people younger and older than 65 years is the most different variable between the TBIMS national database and the national population of adults receiving inpatient TBI rehabilitation.[9,12] Despite the substantial influence of these 2 variables, it is clear that improved precision in the estimation process is demonstrated when all variables are used in the raking process; thus, it is recommended that all known population parameters be used when