International Journal of Medical Informatics 94 (2016) 49–58

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.ijmijournal.com

Assessment of economic status in trauma registries: A new algorithm for generating population-specific clustering-based models of economic status for time-constrained low-resource settings

Lauren Eyler[b,*], Alan Hubbard[a], Catherine Juillard[b]

[a] University of California, Berkeley, School of Public Health, Division of Biostatistics, 50 University Hall #7360, Berkeley, CA 94720-7360, USA
[b] University of California, San Francisco, Center for Global Surgical Studies, San Francisco General Hospital, Box 0807, San Francisco, CA 94143-0807, USA

ARTICLE INFO

Article history:
Received 4 August 2015
Received in revised form 12 April 2016
Accepted 24 May 2016

Keywords:
Economic status
Trauma registries
LMICs
Cluster analysis
Health disparities

ABSTRACT

Objectives: Low and middle-income countries (LMICs) and the world's poor bear a disproportionate share of the global burden of injury. Data regarding disparities in injury are vital to inform injury prevention and trauma systems strengthening interventions targeted towards vulnerable populations, but are limited in LMICs. We aim to facilitate injury disparities research by generating a standardized methodology for assessing economic status in resource-limited country trauma registries where complex metrics such as income, expenditures, and wealth index are infeasible to assess.
Methods: To address this need, we developed a cluster analysis-based algorithm for generating simple population-specific metrics of economic status using nationally representative Demographic and Health Surveys (DHS) household assets data. For a limited number of variables, *g*, our algorithm performs weighted k-medoids clustering of the population using all combinations of *g* asset variables and selects the combination of variables and number of clusters that maximize average silhouette width (ASW).
Results: In simulated datasets containing both randomly distributed variables and "true" population clusters defined by correlated categorical variables, the algorithm selected the correct variable combination and appropriate cluster numbers unless variable correlation was very weak. When used with 2011 Cameroonian DHS data, our algorithm identified twenty economic clusters with ASW 0.80, indicating well-defined population clusters.
Conclusions: This economic model for assessing health disparities will be used in the new Cameroonian six-hospital centralized trauma registry. By describing our standardized methodology and algorithm for generating economic clustering models, we aim to facilitate measurement of health disparities in other trauma registries in resource-limited countries.

1. Introduction

Injury is a significant and growing public health concern, particularly in low- and middle-income countries (LMICs). The 2010 Global Burden of Disease study found that injury accounted for 11% of global Disability Adjusted Life Years (DALYs), which is greater than the total DALYs caused by HIV/AIDS (3.3%), tuberculosis (2.0%), and malaria (3.3%) combined [1]. This burden of injury is distributed inequitably; strong associations between low socioeconomic status and increased rates of morbidity and/or mortality from injuries have been reported among children in England and Wales [2], children in Sweden [3], adult men in multiple European Nations [4], multiple populations in the United States [5], and road traffic injury patients in Cameroon [6]. From the Millennium Development Goal (MDG) initiative, we have learned that development goals cannot be reached unless policies are targeted to address the needs of poor and disadvantaged populations; however, the data necessary to assess health disparities are severely lacking, especially in sub-Saharan Africa [7]. As one of the proposed post-2015 Sustainable Development Goals is to decrease road traffic injuries and fatalities by half by the year 2020 [8], the international health and development communities will need improved data sources for measuring disparities in injury around the world.

Trauma registries in low-resource settings could play an important role in providing this needed data, if they measured patient

Abbreviations: LMICs, low- and middle-income countries; DHS, demographic and health surveys; ASW, average silhouette width.

* Corresponding author.
E-mail addresses: lauren.eyler@ucsf.edu (L. Eyler), hubbard@berkeley.edu (A. Hubbard), catherine.juillard@ucsf.edu (C. Juillard).

http://dx.doi.org/10.1016/j.ijmedinf.2016.05.004
1386-5056/

economic status. Of the 47 trauma registries in LMICs identified by O'Reilly et al. [9], thirteen assess occupation and three assess education, but none include a defined metric of economic status. Because education, occupation, and economic status each affect health outcomes independently and in different ways [10], education and occupation should not be considered as proxies of economic status [11,12]. A brief review of methods for assessing economic status in the economics and development literature demonstrates why LMIC trauma registries have so far failed to measure this important factor in injury risk.

The three most commonly used proxies for economic status are income, expenditures, and assets-based models. Both income and expenditures have been widely criticized as inappropriate for research in resource-limited countries because the distribution of household income and patterns of consumption may vary greatly between household members [13], household survey data on income and expenditures has been found to be highly susceptible to measurement error and bias [14], and assessment of income and consumption rely on lengthy questionnaires [15]. Currently, a more commonly used metric is the DHS' wealth index model, which involves running a principal components analysis (PCA) or factor analysis on DHS variables describing household assets and infrastructure [12]. The first principal component is taken to be the wealth index. Although Juillard et al. [16] were able to collect the data necessary to develop a full wealth index for their pilot trauma registry at Central Hospital of Yaoundé in Cameroon, the time spent with each patient to collect relevant data would not be feasible for a permanent registry.

Several researchers in non-trauma settings have attempted to develop metrics for use in such urgent situations. Pitchforth, van Teijlingen, Graham, and Fitzmaurice [17] and Wilunda et al. [18] viewed the distribution of DHS assets within their catchment populations, selected the four to five asset variables that appeared to best distinguish wealth quintiles, and summed the number of these assets owned to create proxy wealth indices. This common sense approach is practically appealing but difficult to standardize. Tajik and Majdzadeh [19] used regression with stepwise selection of wealth index on individual asset variables to create a simplified linear model consisting of only six variables that best predicted wealth index. However, their misclassification error rate compared to the full wealth index was 35% [19]. Furthermore, all index-based models of wealth force an assumption of linearity on the relationships between the variables, which likely interact in a more multidimensional manner [20]. Patel, Prabhu, Dibley, and Kulkarni [21] used participatory methods to develop a wealth scale. Such consensus-based methods are challenging to implement in larger populations. These various methodological suggestions for creating simplified assets-based wealth scales demonstrate both the common need for a simpler metric of economic status and the lack of a generally accepted methodology for creating one.

Another proposed method for generating models of economic status involves unsupervised cluster analysis. As discussed by Grusky and Weeden [20], advantages of these clustering models are that they allow multidimensional relationships between variables, describe differing groups of poor people with shared economic characteristics rather than assuming a continuous ranking of individuals in society, and are based on sociological theories of the underlying economic class structure of populations. Researchers in the field of agricultural economics have used a variety of clustering algorithms to identify groups with different economic profiles and to target agricultural interventions to those groups who might benefit most [22–24]. To our knowledge, such cluster models have yet to be employed in the analysis of economic status in emergency healthcare settings.

In order to address the need for a standardized method of assessing disparities in injury, we developed a cluster analysis-based algorithm for generating simple, population-specific models of economic status using limited numbers of DHS assets variables. Using DHS data will allow trauma registries to track the economic status of injured patients and periodically compare it to that of the general population in order to assess health systems' abilities to provide care for the least wealthy members of society over time. Potential interventions could then be evaluated in part by assessing the association of their implementation with changes in the economic status of trauma patients being seen at the hospital level. By proposing a standardized and easily replicable methodology, we hope to facilitate and encourage measurement of economic status in trauma registries and other emergency settings around the world.

2. Materials and methods

2.1. Algorithm development

We developed our economic clusters algorithm for use with the Demographic and Health Surveys asset variables that the DHS includes in the wealth index [12]. Some asset variables are binary, indicating household ownership of goods and resources such as agricultural land, televisions, and refrigerators. Other asset variables are nominal categorical variables with up to m levels, such as type of flooring, water supply, or toilet facilities. Each observation O_i is one household's set of responses to these t variables with an inverse probability weight w assigned based on the nature of the survey design.

$$O^n = (O_1, \ldots, O_n)$$

$$\text{where} O = (X_1, X_2, \ldots, X_t, w) \text{and}$$

$$X_i \in \left\{1, 2, 3, \ldots m_i\right\}$$

We are assuming a model of economic status in which there are underlying clusters of individuals within the population that share similar asset profiles. Cluster-based models are advantageous for socioeconomic assessment because they require few statistical assumptions about the underlying distributions of asset variables within the population compared to linear index-based models and because they are consistent with class-based sociological theories of societal economic structure [20].

As recommended by Hennig and Liao [25] for dissimilarity-based cluster analysis of economic classes, we use unsupervised k-medoids clustering to define population clusters. The k-medoids clustering algorithm iteratively assigns individual observations O_i as cluster "medoids" $O^*{}_j$. Medoids function as the center or most representative data point for each cluster. The algorithm assigns all other observations to clusters based on how similar they are to the cluster medoids. Dissimilarity or distance of each observation from each medoid is denoted by $D_{i,j}$. The algorithm determines the set of k medoids and corresponding cluster membership of all observations that minimizes the sum of the distances of all individual observations to their cluster medoids. As described by Hennig and Liao [25], the target function which the k-medoids clustering algorithm seeks to minimize over all configurations of $O^*{}_j$ is

$$Sum\,Dist = \sum_{i=1}^{n} \underset{j \in \{1,\ldots,k\}}{\operatorname{argmin}} D\left(O_i, O_j^*\right)$$

We conducted this analysis using R version 0.98.1103 [26]. We created our dissimilarity matrix D using R function *daisy* from package *cluster* [27] with Gower's dissimilarity metric for mixed type variables [28]. For the symmetric binary and nominal categorical variables used in this study, the dissimilarity or distance between two individuals is simply the proportion of variables for which