UNIVERSITY of York

HEDG

HEALTH, ECONOMETRICS AND DATA GROUP

WP 19/22

A latent class approach to inequity in health using biomarker data

Vincenzo Carrieri; Apostolos Davillas & Andrew M Jones

October 2019

http://www.york.ac.uk/economics/postgrad/herc/hedg/wps/

A latent class approach to inequity in health using biomarker data

Vincenzo Carrieri¶
Apostolos Davillas†
Andrew M Jones‡

Abstract

We develop an empirical approach to analyse, measure and decompose Inequality of Opportunity (IOp) in health, based on a latent class model. This addresses some of the limitations that affect earlier work in this literature concerning the definition of types, such as partial observability, the *ad hoc* selection of circumstances, the curse of dimensionality and unobserved type-specific heterogeneity that may lead to either upwardly or downwardly biased estimates of IOp. We apply the latent class approach to measure IOp in allostatic load, a composite measure of our biomarker data. Using data from *Understanding Society* (UKHLS), we find that a latent class model with three latent types best fits the data and that these types differ in terms of their observed circumstances. Decomposition analysis shows that about two-thirds of the total inequality in allostatic load can be attributed to the direct and indirect contribution of circumstances.

Keywords: equality of opportunity; health equity; biomarkers; finite mixture models; latent class models; decomposition analysis

JEL codes: C1, D63, I12, I14.

Understanding Society is an initiative funded by the Economic and Social Research Council and various Government Departments, with scientific leadership by the Institute for Social and Economic Research, University of Essex, and survey delivery by NatCen Social Research and Kantar Public. The research data are distributed by the UK Data Service. Andrew Jones acknowledges funding from the Leverhulme Trust Major Research Fellowship (MRF-2016-004). The funders, data creators and UK Data Service have no responsibility for the contents of this paper.
¶ Department of Law, Economics and Sociology, "Magna Graecia" University of Catanzaro; RWI-Research Network; Department of Economics, Lancaster University Management School
† Office of Health Economics (OHE), London and Institute for Social and Economic Research, University of Essex.
‡ Department of Economics and Related Studies, University of York; Centre for Health Economics, Monash University

1 Introduction

Based on Roemer's (1998, 2002) influential formalisation of the concept, a large body of empirical research has dealt with the assessment of inequality of opportunity (IOp) for a variety of measures of well-being. The IOp literature argues that the egalitarian framework does not necessarily indicate equality of the distribution of outcomes *per se* but emphasises the role of individual responsibility in defining a "fair" distribution. Early contributions to the IOp literature have focused mainly on income (see Ramos and van de Gaer (2016) and Roemer and Trannoy (2016) for reviews). More recently, a growing literature has addressed the measurement of IOp in other relevant dimensions of individual well-being such as education (Ferreira and Gignoux, 2013) and health (e.g., Rosa Dias, 2009; Rosa Dias, 2010; Trannoy et al., 2010; Jusot, et al., 2013; García-Gómez et al., 2015; Deutsch et al., 2018).

This literature separates the factors associated with an outcome of interest into two components: 'circumstances', which are not under individual responsibility and are viewed as an illegitimate or unfair source of inequality, and 'efforts' for which, to some extent, individuals are held responsible and that are viewed as a legitimate source of inequality. Following Roemer (1998, 2002), the IOp literature often defines types as a group of individuals who share the same set of circumstances, such as parental background and early life circumstances, emphasising the intergenerational transmission of disadvantage (e.g., Aaberge et al., 2011; Carrieri and Jones, 2018; Fleurbaey and Peragine, 2013; Ramos and van de Gaer, 2016; Trannoy et al., 2010). In the context of health equity, Fleurbaey and Schokkaert (2009, 2012) take a broader perspective that uses the responsibility cut to distinguish factors that are seen as fair sources of inequality of outcomes and those that are seen as unfair, with the health variations attributed to the latter is regarded as health inequity. In this study, we adopt a social perspective and draw on the socio-legal context of the UK health system to define the sources of the unfair variation. Thus, for example, following the Equality Act 2010 and how the public sector equality duty is translated into NHS guidelines for reducing health inequalities (e.g., NHS England, 2017), we include age and sex among our set of circumstances variables along with measures of parental socioeconomic status.

A key empirical challenge in these analyses is the definition of types. It is difficult to devise a criterion to make the Roemer model operational, especially because the original model does not provide practical guidance for neither the number nor the combination of circumstances that should be used to define social types (Li Donni et al., 2015). This implies that a large part of the existing empirical research in IOp may have a number of limitations. First, researchers may observe only a limited set of circumstances, with the partial observability of the circumstances is often a common feature of IOp studies (see Brunori et al., 2019 and Li Donni et al., 2015 for a relevant discussion). This may lead to an underestimation of the share of illegitimate inequality. Second, researchers often rely on *ad hoc* definitions of types according to exposure to a small number of circumstances which, although they may be guided by the norms and conventions of the society being analysed, are more or less arbitrarily selected by the researcher (Li Donni et al., 2015). Third, the combination of selected circumstances into types may result in a trade-off between the number of types and the sample size for each type. For example, the high correlation between different measures of parental socioeconomic status can make it hard to define clear cut and mutually exclusive categories, resulting in types with few observations, which may lead to overestimates in the measurement of IOp (Brunori et al., 2019). Researchers often address these problems by using a limited number of circumstances or an arbitrary aggregation of socioeconomic categories. The curse of dimensionality may imply severe limitations given that stochastic dominance tests, often employed as a first stage to identify the presence of IOp, are highly sensitive to the choice of circumstance variables, as are results from analyses that involve separate regressions by type (e.g., Bourguignon and Ferreira, 2007; Carrieri and Jones, 2018; Garcia-Gomez et al., 2015). Beyond nonparametric analysis, reliability of parametric IOp estimates may also require a sufficient number of observations in each category to characterize circumstances (Brunori et al., 2019).

Building on the work of Bago d'Uva et al. (2009) on horizontal inequity in health care, Balia and Jones (2011) on IOp in mortality, and Li Donni et al. (2015) on IOp in life satisfaction, we propose an empirical approach to quantify and decompose IOp in health based on latent class models. We use data from *Understanding Society: the UK Household Longitudinal Study (UKHLS)*, a nationally representative study that allows for objectively measured nurse-collected and blood-based biomarker data, and for a rich set of circumstance and effort variables. Specifically, we apply finite mixture models (FMMs), a semi-parametric approach to model unobserved heterogeneity regarding type membership, which, unlike most of the existing IOp studies, avoids *a-priori* grouping of individuals into types. Instead, FFMs are a semiparametric method to classify individuals into latent classes (types), with the likelihood of latent class membership to be a function of the set of observed circumstance variables. This analysis allows us to select the optimal number of latent classes (types) that are consistent with the data generation process.

A potential disadvantage of defining social types in terms of latent classes is that they are treated as a "black box", which may be hard to interpret and to assign a normative significance. We therefore augment our FMM analysis with post-estimation analysis to help characterise the latent types in terms of the combination of observed circumstances that each of them may reflect, and classify individuals into the different latent types based on the estimated posterior probabilities of class membership.

Capitalising on this useful feature of FMMs to classify individuals into latent types, we further contribute to the literature by adapting and extending a recently developed decomposition technique to decompose health inequality (Carrieri and Jones, 2018). This analysis allows us to decompose total inequality in health into the direct contribution of circumstances, their indirect contribution via the heterogenous association of efforts with health by type and the direct contribution of efforts themselves. Extending our FMM analysis to decompose health inequality and identify the role of IOp offers a number of advantages compared to earlier work in this literature concerning the definition of types. Our analysis allows the optimal number of types and the particular combination of circumstances that are used to define each type to be determined by our model and reflect the data generation process. This avoids arbitrary combinations of circumstance variables to define types or the use of an excessive number of types that may impose upward bias in the IOp measurement (Brunori et al., 2019). The FMM methodology is also helpful here since it accounts for unobserved type-specific heterogeneity in the sense of exploring differences in the association between efforts and the health outcome by latent type. Dealing with unobserved heterogeneity regarding type membership and simultaneously allowing for heterogeneous effort-health outcome associations by types is of critical importance measuring IOp and better understanding its underlying sources.

Finally, this paper further contributes to the health equity literature by being one of the few studies that is not based on self-reported measures to proxy individual health.[1] We use a composite biological measure that captures several health dimensions, spanning adiposity, blood pressure, inflammation, blood sugar levels and cholesterol levels. Similar measures are used to capture allostatic load and are considered as measures of "wear and tear" of the body that accumulates as individuals are exposed to chronic psychosocial stressors (Davillas and Pudney, 2017; Howard and Sparks, 2016; Seeman et al., 2004). As such, allostatic load is an ideal health measure for the purpose of the measurement of IOp because it captures physiological responses that are associated with stress and the process through which economic and social circumstances may get "under the skin" across the lifespan (McEwen, 2015; Seeman et al., 2004).

[1] Self-assessed health (SAH), one of the most popular self-reported health measures, is an inherently categorical and ordinal measure and may be subject to misreporting and is associated with comparability problems at both the individual level and between countries (eg., Bago d'Uva et al. 2008). This reporting bias has been shown to vary systematically with a number of socioeconomic characteristics that are often used to explore health inequalities, which raises doubts about the robustness of studies based on self-reported health indicators (e.g., Bago d'Uva et al., 2009; Crossley and Kennedy, 2002). More fundamentally, the ordinal scaling of SAH limits the range of inequality indices that can be used as many of these require cardinal outcomes. Recent work by Bond and Lang (2019) highlights the sensitivity of conclusions drawn from ordinal data to the scaling imposed on it.

We find that a latent class model with three unobserved types provides the best fit with our data. The profiles of these types can be characterised in terms of differences in their observed demographic and parental circumstances. After classifying individuals into classes using modal assignments, decomposition analysis shows that about 50% of the total inequality in our composite health measure (allostatic load) is attributed to the direct contribution of demographic and parental circumstances. Circumstances exert an indirect contribution to the total inequalities of around 14%, though differences in the association between our effort variables and allostatic load across types. However, the direct contribution of efforts is less important, having a contribution of around 3%.

2 Methods

Following the seminal work of Roemer (1998, 2002), the IOp literature assumes a responsibility cut by which factors associated with individual attainments can be grouped into two categories: a) effort factors, for which individuals should be held partially responsible, and b) circumstances which are beyond individuals' control. In the case of health, following the IOp literature (e.g., Carrieri and Jones, 2018, Jusot et al., 2013, Rosa Dias, 2010), a generalised health production function for individual health outcomes (y_i) can be defined as a function of a vector of circumstances c_i and of efforts e_i. Assuming that circumstances are not affected by efforts, while efforts may be influenced by circumstances (Roemer, 1998, 2002), we can write:

$$y_i = h(c_i, e(c_i, v_i), u_i) \tag{1}$$

where v_i and u_i are unobserved error terms which capture the random variation in the realised outcomes. This reflects the fact that observed realisations of health outcomes are inherently random, sometimes labelled as 'luck' in the IOp literature (Lefranc et al., 2009; Lefranc and Trannoy, 2017). To be specific, v_i represents random variation in effort that is independent of c and u_i represents random variation in the outcome that is independent of c and e. The latent class specification is used to model the conditional density function $f(y_i|\boldsymbol{c_i}, \boldsymbol{e_i})$ that is implied by equation (1). A fundamental feature of the Roemer approach is the fact that the distribution of effort within each type is itself a characteristic of that type and, since this is assumed to be beyond individual responsibility, it constitutes a circumstance itself. This implies that, in addition to assuming a partitioning between c and e, the IOp model assumes that effort is a function of circumstances, with circumstances being pre-determined. Effort, therefore, mediates the relationship between circumstances and outcomes, and it is meaningful to consider the direct and the indirect contribution of circumstances to the inequality in outcomes. One of the strengths of our FMM analysis and our decomposition analysis, is that it allows us to explore the type-specific unobserved heterogeneity in the association between our health measure and efforts and identify the direct and indirect role of circumstances on shaping inequalities in our health outcome.

In this context, researchers interested in quantifying IOp in well-being outcomes (including health), typically define social types, i.e., groups of individuals who share exposure to the same circumstances, and then measure IOp between these types (e.g., Aaberge et al., 2011; Carrieri and Jones, 2018; Fleurbaey and Peragine, 2013; Ramos and Van de Gaer, 2016; Trannoy et al., 2010). Roemer (2002) defines social types consisting of individuals who share exposure to the same set of circumstances. Although the theoretical framework for the concept of types is well developed, implementation in applied work is less straightforward. As discussed in the introduction, types are often defined in an *ad hoc* way in empirical work and they are partially observable to researchers (Li Donni et al., 2015).

In this context, latent class or finite mixture models (FMMs) offer a number of important advantages (Bago d'Uva and Jones, 2009; Bago d'Uva et al., 2009; Balia and Jones, 2011; Li Donni et al, 2015). FMMs provide an intuitive representation of unobserved heterogeneity that may exist in the data in a parsimonious number of latent classes. The prior probabilities of membership of these classes can be parameterized to depend on observed circumstance variables, interpreting the latent classes as unobserved types in the context of the IOp framework. Additionally, FMMs are particularly flexible because they do not require the researcher to assume, *ex-ante*, the number of latent classes, nor to provide any *a-priori* grouping based on the observed circumstances. Another advantage of FMMs is that they are semiparametric and do not require distributional assumptions for the mixing variable.

In the FMM, the conditional density of our health outcome variable, allostatic load, is assumed to be drawn from a population characterised as an additive mixture of K (j=1,…,K) distinct classes in proportions of ρ_j, where, $0 \leq \rho_j \leq 1$, $\sum_{j=1}^{K} \rho_j = 1$. The concept of IOp is based on the notion that types are defined on the basis of individual's exposure to circumstance variables (Roemer, 1998). The mixture probabilities of class membership are assumed to be a function of the set of observed circumstance variables ($\boldsymbol{c_i}$):

$$f(y_i|\boldsymbol{c_i}, \boldsymbol{e_i}) = \sum_{j=1}^{K} \rho_j(\boldsymbol{c_i}, \boldsymbol{\alpha_j}) f_j(y_i|\boldsymbol{e_i}, \kappa_j, \boldsymbol{\beta_j}, \boldsymbol{\theta_j}) \quad (2)$$

where, $f_j(.)$ is the j[th] density, $\boldsymbol{\theta_j}$ stands for the vector of parameters describing the density function f_j, ρ_j are the mixture probabilities and $\boldsymbol{e_i}$ is the vector of effort variables, $\boldsymbol{\alpha_j}$ is a set of circumstance coefficients to be estimated from a multinomial logit model for class membership. $\boldsymbol{\beta_j}$ is the set of effort coefficients for each type j and κ_j are the relevant constant terms. Estimation of equation (2) also allows us to explore the heterogeneous association between efforts and allostatic load across the different latent types.

In FMMs, the prior probability for the j[th] latent class can be expressed as a function of circumstance variables using a multinomial logit transformation. Typically, FMMs can allow for large set of discrete and continuous densities ($f_j(.)$) of different types. For our analysis, we estimate FMMs assuming that the outcome variable (allostatic load) is a mixture of a number of normal distributions, each with its own mean and variance. The normal provides a good fit for our measure of allostatic load.

The choice of the appropriate number of latent types (K) is crucial for FMMs; we use statistical information criteria to identify the FMM with the number of classes that makes the best statistical fit (Cameron and Trivedi, 2010). A caveat for the use of FMMs is the risk that outliers in the data may be captured by additional mixture components. Hence, it is desirable that FMM estimation results in latent classes that account for a sufficient number of observations as well as having meaningful posterior differences in outcomes across the different latent classes (Cameron and Trivedi, 2010; Deb et al., 2011). We consider these matters as additional criteria for selection of the appropriate number of classes.

Once the number of latent classes (types), K, is selected we can use the parameter estimates from the model to calculate the posterior probability of each individual being assigned to a given latent class $j = 1,2 \ldots K$. The posterior probability of membership in each latent class (type) is calculated conditional on all c, e and the outcome variable (y) as:

$$\Pr(y_i \in \nu | \, \boldsymbol{\theta}, y_i) = \frac{\rho_\nu f_\nu(y_i|\boldsymbol{\theta}_\nu)}{\sum_{j=1}^{K} \rho_j f_j(y_i|\boldsymbol{\theta}_j)}, \ \forall \ \nu = 1,2, \ldots K. \tag{3}$$

For each individual (i), K posterior probabilities are estimated, one for the membership of each type. Following the common practice in the literature (e.g., Deb et al., 2011; Li Donni et al., 2015), we assign each individual to the type with the highest posterior probability (known as *modal assignment*). It has been shown that this modal assignment is problematic when, for a substantial number of individuals, the highest and the next-highest posterior probabilities of belonging to two or more different types are particularly close (e.g., Vermunt and Magidson, 2004). However, as we will show later in the paper, this is not an issue in our empirical analysis. Our FMMs are estimated using maximum likelihood techniques, with robust standard errors.

2.1 Decomposition of IOp by factor components

To explore the contribution of circumstances and efforts in shaping total inequalities in health, we adapt and extend a recently proposed decomposition approach for IOp

(Carrieri and Jones, 2018) combined with our FMM analysis. The FMMs allow us to classify individuals into types, taking into account the complex circumstance profile that may be reflected in each different type. Modal assignments, based on the posterior probabilities from the latent class analysis, facilitate classification of individuals into each of the different latent types (j = 1,2,3). FMM also allows us to capture unobserved heterogeneity in the sense of individual differences in the association between efforts and allostatic load across different types.

We have used the variance as our inequality measure, given the fact that recent contributions to the IOp literature have favoured the variance as an absolute measure of health inequality (see e.g., Fleurbaey and Schokkaert, 2009, 2012; Jusot, et al., 2013; Carrieri and Jones, 2018). Specifically, we have used an extension to the variance decomposition of Shorrocks (1982), allowing us to decompose the role of circumstance and efforts on shaping inequalities in health (Carrieri and Jones, 2018).

To recap, κ_j in equation (2) reflects differences in allostatic load across types and β_j the heterogeneous effort-allostatic load associations by type. To provide a benchmark for our decomposition analysis, we first define the effort of each individual given their type (B_i), as the product of the effort variable and the associated heterogeneous slope coefficients ($\boldsymbol{\beta_j}$) by type obtained from the FMM (equation 2):

$$B_i = \boldsymbol{\beta_j}\mathbf{e_i},\ i \in j \tag{4}$$

Then, as a benchmark, we use the following weighted averages within and across types after assigning individuals into types using modal assignments, where π_j denotes the estimated share of each type:

$$\bar{y} = \sum_j \pi_j \bar{y}_j\ ;\ \bar{B} = \sum_j \pi_j \bar{B}_j\ ;\ \bar{\kappa} = \sum_j \pi_j \kappa_j\ ,\bar{\beta} = \sum_j \pi_j \boldsymbol{\beta_j} \tag{5}$$

As shown in Carrieri and Jones (2018), applying the Shorrocks decomposition gives:

$$\mathrm{Var}(y) = \mathrm{cov}(\kappa_j - \bar{\kappa}, y) + \mathrm{cov}\left((\bar{B}_j - \bar{B}), y\right) + \mathrm{cov}\left((B_i - \bar{B}_j), y\right) + \mathrm{cov}(u_i^j, y) \tag{6}$$

The first term in equation (6) is the contribution of the variation of the intercepts κ_j across types (equation 2), centred at the pooled mean. This term measures the direct contribution of circumstances to the overall inequality. The second term reflects the indirect contribution of circumstances to overall inequality, capturing variation in the average level of effort within each type around the pooled mean of effort. The third term is the contribution of the within-type variation in effort to the overall health inequality. In normative terms, this represents the contribution of effort. The final term measures the contribution of residual factors u to overall inequality.

3 Data

The data come from UKHLS, a longitudinal, nationally representative study of the UK. In this study, we use the General Population Sample (GPS) component of UKHLS, a random sample of the general population. As part of wave 2 (2010-2011), nurse-measured and non-fasted blood-based biomarkers were collected for the GPS.[2] Exclusion of missing data on our biomarkers, circumstance and effort variables results in a working sample of 6,111 individuals.

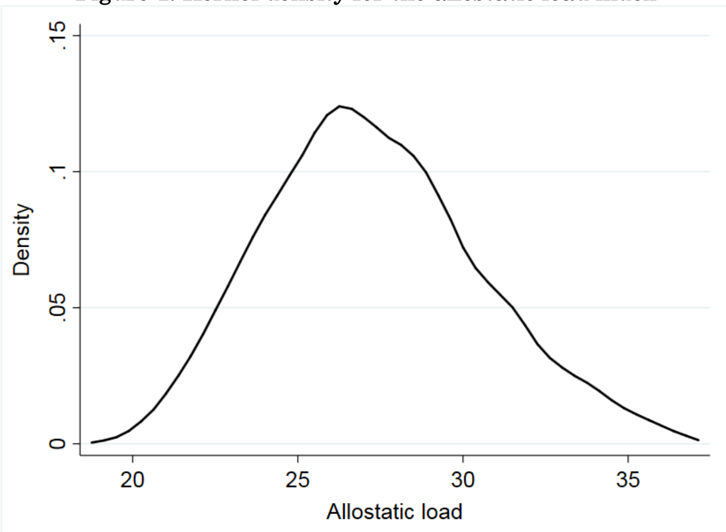
A multi-system biological risk measure: allostatic load

We use a cumulative biomarker index often called allostatic load (e.g., Davillas and Pudney, 2017; Howard and Sparks, 2016; Seeman et al., 2004). The allostatic load is regarded as a biological risk score reflecting the cumulative effects of chronic exposure to psychosocial and environmental challenges or stressors that may leads to significant physiological dysregulation and increased morbidity and mortality risks (Howard and Sparks, 2016; Seeman et al., 2004). As such, allostatic load is of particular relevance in our analysis as IOp is based on concerns about a lasting effect of circumstances on individuals' long-term health.

Our index combines biomarkers for adiposity, blood pressure, inflammation, blood sugar levels and cholesterol (see Table A1, appendix for a description of the relevant biomarkers). Each of these biomarkers is transformed into standard deviation units and then summed to define allostatic load. It has been shown that a single measure of the different biomarkers is sufficient to measure allostatic load (Howard and Sparks, 2016). Higher values of allostatic load indicate worse health. Given that allostatic load is modelled here as a mixture of normals, it is notable that the density of allostatic load is unimodal and fairly symmetric (Figure 1).

[2] Respondents were eligible for nurse visits if they were aged 16+, lived in England, Wales, or Scotland, and were not pregnant. Blood sample collections were further restricted to those who had no clotting disorders and no history of fits.

Figure 1. Kernel density for the allostatic load index



Circumstances

Our set of circumstance variables embodies the ethical position of the responsibility cut, defining illegitimate sources of health inequality. For the choice of circumstance variables, we follow the recent literature on health equity along with the UK policy and legal context (Davillas and Jones, 2018; Carrieri and Jones, 2018; Rosa Dias, 2009, 2010; Jusot et al., 2013).

Drawing on the socio-legal context in the UK, we treat sex and age as circumstances; sex and age are considered as protected characteristics under the Equality Act of 2010[3]. Specifically, in this study we account for gender, age as a continuous variable and their interaction variable. Socioeconomic status (SES) in childhood has been an important concern of the existing literature on IOp. Childhood SES is regarded as an important source of IOp in health, being beyond individual's control and exerting a lasting effect on individual's adult health (Jusot et al., 2013; Rosa Dias, 2009, 2010). In our analysis we use both parental occupational status and parental education to proxy childhood SES. Two categorical variables (one for each parent) are used to capture the occupational status of the respondent's mother and father, when the respondent was aged 14: not working (reference category), four occupation skill levels and a category for missing data. To construct these variables the occupational skill levels are based on the skill level structure of the Standard Occupational Classification 2010. Given the high correlation between mother's and father's education, we combine them creating a measure capturing the highest parental education level (Kenkel et al., 2006). This is a five-category variable measured as: left school with no/some qualification (reference category), post-school

[3] For example, NHS England suggests actions to advance equality of opportunity in health, particularly relevant to patient's age and gender, characteristics that are "protected" under the Equality Act (NHS England, 2017).

qualification/certificate (e.g., an apprenticeship), degree (university or other higher-education degree) and a missing data category.

Efforts

In the concept of IOp in health, effort variables typically indicate decisions to invest in health capital, such as health-related lifestyles (e.g., Balia and Jones, 2011; Carrieri and Jones, 2018, Rosa Dias, 2010). Smoking status is captured by a categorical variable: current smoker, ex-smoker and never-smoker (reference category). Unhealthy dietary habits are captured by a dummy taking the value of one when the individual does not comply with the recommendation of five portions of fruits or vegetables per day and zero otherwise and an indicator for usual consumption of white (versus non-white) bread. Physical inactivity is captured by a dummy for not being a frequent walker (walk less than 5 times per week) and by a categorical variable for the frequency of sports participation: 3+ times/week (reference category), 1-3 times/week, once per month or not at all.

4 Results

Table 1 presents the values of the AIC and BIC for each FMM estimated with different numbers of types.[4] The model with three latent classes is the one that minimises the BIC and, thus, selected as our preferable model here (Cameron and Trivedi, 2010). Although FMMs with a higher number of latent classes have lower AIC values, BIC values are increasing compared to the FMM with three latent classes. Further support for the FMM with three classes comes from the fact that it results in reliably differentiated latent classes (Table 2); each latent class accounts for a sufficient number of observations and the mean values of allostatic load (our outcome variables) are distinct across the three latent classes (there is no overlap in their confidence intervals). For example, Table A2 (Appendix) shows that the FMMs with a higher number of latent classes (especially those with five classes and above), have one or more latent classes that account for a fairly small part of the population or are characterised by non-distinctive latent classes with respect to the predicted allostatic load across types.

Table 1. FMMs for allostatic load: AIC and BIC.

Number of latent classes (types)	AIC	BIC
K=1	31292	31359
K=2	29878	30126
K=3	29625	30055
K=4	29592	30203
K=5	29553	30340
K=6	29561	30535
K=7	29521	30671

[4] Table A2 (Appendix) shows the corresponding full set of posterior probabilities and mean allostatic load values by latent class.

Focusing on our preferred FMM with three latent types (Table 2), we find that about 21% of our sample is estimated to belong to type 1 (the latent class with the lowest health risk, i.e., with the lowest mean allostatic load value), 40% in type 2 (the latent class with the second-lowest allostatic load) and 39% in type 3 (the type with the highest allostatic load).

Table 2. Latent class (types) probabilities and predicted mean allostatic load: FMM with three latent types.

	Latent class probabilities ρ_j (%)	Predicted mean allostatic load
Type 1	20.63 (17.10; 24.67)	23.50 (23.16; 23.84)
Type 2	40.35 (33.20; 47.94)	26.63 (26.23; 27.04)
Type 3	39.02 (29.61; 49.32)	29.89 (29.35; 30.43)

Note: 95% confidence intervals in parenthesis

4.1 Modal assignment of individuals to the latent types

Using the parameter estimates for our preferred 3-class FMM, we estimate posterior probabilities of membership on each of the three latent types for each individual. Modal assignments indicate that individuals are assigned to the type with the highest posterior probability. To explore this issue further, Table 3 presents the mean values of the posterior probabilities of class membership conditional on modal type assignment. Focusing on those who are classified into type 1 using the modal assignment ("Type 1" column in Table 3), we find that the mean posterior probability of belonging to type 1 ($\Pr(y_i \in type\ 1)$) is around 84%, with 90% of those individuals having posterior probabilities to belong to this type (i.e., $\Pr(y_i \in type\ 1)$) of above 57.9% (as shown by the relevant quantile statistics; Q10, Q50 and Q90). The corresponding mean posterior probabilities of belonging to types 2 (around 15%) and 3 (around 1.4%) are much lower. Similarly, the mean posterior probability of belonging to type 2 is 71% for those who are assigned to type 2 using the modal assignment (Table 3, column "Type 2"). Modal assignments to type 3 seem sensible also given the very high mean posterior probability for type 3 membership (around 83%; Table 4, column "Type 3").

Table 3. Posterior latent class (type) membership probabilities conditional on modal assignment of individuals into types: FMM with three latent types.

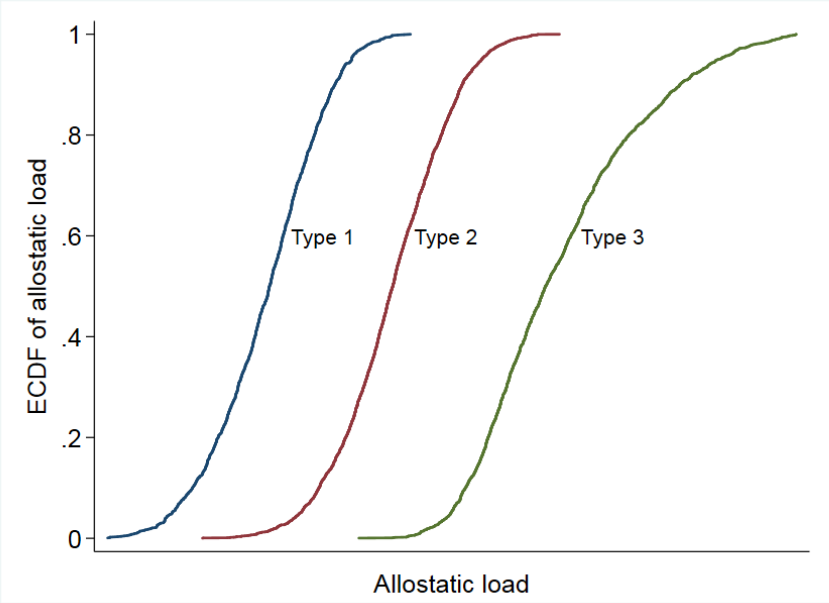
	Modal assignment into latent classes (types)		
Posterior probabilities	Type 1	Type 2	Type 3
$\mathbf{Pr}(\boldsymbol{y_i} \in \boldsymbol{type\ 1})$			
Mean (i)	**83.6%**	7.6%	0.0%
Q10	57.9%	0.2%	0.0%
Q90	99.3%	27.9%	0.1%
$\mathbf{Pr}(\boldsymbol{y_i} \in \boldsymbol{type\ 2})$			
Mean (ii)	15.0%	**70.8%**	16.9%
Q10	0.6%	54.7%	1.2%
Q90	40.6%	85.0%	43.2%
$\mathbf{Pr}(\boldsymbol{y_i} \in \boldsymbol{type\ 3})$			
Mean (iii)	1.4%	21.6%	**83.1%**
Q10	0.2%	6.1%	56.7%
Q90	3.7%	41.0%	99.9%
Total (sum of rows i, ii, iii)	100%	100%	100%

Notes: Q10 and Q90 stand for the 10th and 90th quantiles of the posterior probabilities conditional on modal assignment of individuals into types.

Overall, these results show that modal assignments across the three types are sensible in our analysis. For the vast majority of individuals, there are clear differences between the highest posterior probability of belonging to a certain latent type and other two posterior probabilities for the remaining types.

Figure 2 presents the graphical illustration of the empirical distribution functions for allostatic load by types, defined using the modal assignment. The graph shows a clear difference in the distribution of allostatic load across types confirming our results in Table 2. From an IOp perspective, these distributions can be interpreted as representing the opportunity sets facing each of the types, in terms of the distribution of health outcomes available to them, bearing in mind that a higher score of allostatic load implies worse health. There appears to be first order stochastic dominance across the three types. The contrast between the distributions for types 1 and 3 is particularly striking with the non-overlapping support for the two distributions suggesting zero order stochastic dominance.

Figure 2. Allostatic load distributions by types (defined using the modal assignment): FMM with three latent types.



4.2. Characterising the profile of the latent types

The analysis so far does not characterise the profile of the three latent types in terms of the observed circumstances. In the concept of IOp, types are defined on the basis of individuals' exposure to circumstance variables and, thus, identifying whether each latent type reflects more or less disadvantaged observed circumstances is of particular importance. Table 4 shows, in each row, the mean posterior probabilities of belonging to each of the three latent types ($\Pr(y_i \in type\ 1, 2\ and\ 3)$) conditional on selected observed circumstances. Since, by construction of the latent class model of type membership, each individual is assumed to belong to a single type, the probabilities in each row always add up to 1. For the case of the categorical circumstance variables, mean posterior probabilities are calculated for the most and least deprived category; given the interaction between our continuous age variable and gender used in our analysis, the relevant mean posterior probabilities for the three latent types are calculated at selected age groups by gender that are defined here for presentation purposes. It should be noted that our set of circumstance variables are jointly highly significant as determinants of individuals class membership in the multinomial logit model for class membership of our FMM.

Table 4. Posterior type membership probabilities, conditional on observed circumstances: FMM with three latent types.

Observed circumstances	*Type 1*	*Type 2*	*Type 3*
Mother's occupational status			
Highest group (skill level 4)	0.416	0.374	0.210
Lowest group (unemployed)	0.130	0.410	0.460
Father's occupational status			
Highest group (skill level 4)	0.309	0.431	0.261
Lowest group (unemployed)	0.182	0.476	0.342
Parental education			
Degree	0.423	0.434	0.143
No qualification	0.152	0.406	0.442
Age-gender profile			
Males 16-30	0.595	0.298	0.107
Males 31-45	0.146	0.571	0.282
Males 46-60	0.008	0.536	0.456
Males 61-75	0.001	0.394	0.605
Males 76+	0.000	0.266	0.734
Females 16-30	0.728	0.216	0.057
Females 31-45	0.518	0.327	0.154
Females 46-60	0.251	0.432	0.317
Females 61-75	0.041	0.401	0.558
Females 76+	0.010	0.301	0.689

Notes: The probabilities in each table's row add up to 1.

Younger individuals (particularly females), those having a mother (and to lesser extent a father) with higher occupational status as well as those with more educated parents are most likely to belong to type 1 (Table 4). For example, the posterior probability to belong to type 1 for an individual who experienced the most advantaged maternal occupational status during childhood is higher (i.e., 0.416) as compared to type 2 (0.374) and type 3 (0.210). The type 2 latent class lies between the least (type 1) and the most deprived (type 3) types. Specifically, although it is more likely to consist of individuals at earlier to later middle ages, those who had a father working in a highly skilled job (skill level 4) and/or at least one parent with a degree qualification, we also observe large posterior probabilities for those at the lowest parental occupation and educational groups to belong to type 2. Type 3 clearly differs from the other two types to the extent that members are more likely to come from those who are older (and males) and from those who experienced the lowest parental occupation and educational status during their childhood; for example, the probability of an individual, who experienced the more deprived parental occupational status and parental education categories, belonging to type 3 versus the other two types, is the largest. For example, the posterior probability for belonging to type 3 for an individual who experienced the most deprived parental education is higher (i.e., 0.422) compared to type 1 (0.152) and type 2 (0.406). Overall, these results reveal a set of three fully characterised latent types, each of which reflects a complex set of observed circumstances. This complex profile of types, obtained using latent class techniques, indicates what may have been missed if single circumstances were chosen to define types.

4.3 Decomposition of overall inequality

The analysis so far shows that modal assignments of individuals into the three latent types are feasible (subsections 4.1 and 4.2). Beyond the definition of types, the FMM analysis also allows us to account for the type-specific heterogeneity in the association between effort variables and allostatic load. Both the latter and the definition of types are of particular importance in our decomposition of inequality analysis.

Specifically, our FMM results show considerable heterogeneity in the association between effort variables and allostatic load (Table A3, Appendix). Overall, all variables reflecting less healthy lifestyles (given the reference categories) show a positive association with higher allostatic load values indicating higher health risks; the associations become more evident in types 2 and 3, which are the types facing more adverse circumstances compared to type 1. A formal statistical test rejects the null hypothesis that the effort coefficients are equal across types (p-value=0.000).

Table 5 presents the results of the decomposition analysis, allowing us to decompose the sources of inequality in allostatic load and the role of IOp on shaping these inequalities based on the results from our FMMs (see sub-section 2.1 for details on the decomposition). The table shows the direct contribution of circumstances, the contribution of efforts as well as the indirect contribution of circumstances via efforts to the overall inequality in allostatic load.

Our results show that our circumstance variables account for most of the total inequality, with the direct contribution of circumstances being the most important component. Specifically, about 50% of the total inequality in our composite health measure is attributed to the direct contribution of circumstances. The contribution of the role of indirect circumstances via efforts show that circumstances exert an indirect contribution to the total inequalities of around 14%, though differences in the association between our effort variables and allostatic load across types. The detailed decomposition of indirect circumstances show that contributions are positive and indicates that the association between the lifestyle variables and allostatic load is larger for the types who have worse health. Lack of frequent physical activity, unhealthy food habits and smoking are the first, second and third most important indirect mechanisms, respectively. Less important however is the direct contribution of the effort variables (within types) in explaining total inequality in allostatic load (accounting for around only 3%).

Table 5. Decomposition of variance in allostatic load based on the FMM with three latent types.

Circumstances and efforts	Absolute contribution	% contribution
Direct circumstances	**5.18**	**49.57%**
Indirect circumstances via efforts		
Smoking†	0.24	2.29%
Non-compliance: 5 fruits/vegetables	0.26	2.48%
White bread	0.12	1.12%
Non-frequent walking†	0.42	4.00%
Sports activity†	0.41	3.89%
Total indirect circumstances via efforts	**1.45**	**13.88%**
Direct efforts		
Smoking†	0.10	0.95%
Non-compliance: 5 fruits/vegetables per day	0.02	0.15%
White bread	0.04	0.41%
Non-frequent walking†	0.03	0.31%
Sports activity†	0.13	1.25%
Total direct efforts	**0.32**	**3.06%**
Residual	**3.77**	**33.49%**
Total Variance	**10.45**	**100%**

†Absolute and percentage contributions represent the total contribution of all the categories of the relevant categorical variables included in our models.

5 Conclusion

A key empirical and practical challenge in all IOp literature is the definition of types. In this paper, we have proposed an empirical approach to both analyse and decompose IOp in a composite biomarker measure, allostatic load. Our analysis addresses some of the limitations that affect earlier work, namely the partial observability, the *ad hoc* selection of circumstances and the curse of dimensionality. We use FMMs, a semi-parametric approach to model unobserved heterogeneity regarding type membership, which avoids *a-priori* grouping of individuals into types. This analysis facilitates selection of the number of latent classes (types) and allows us to characterise the latent types in terms of the combination of observed circumstances that they represent, as well as classifying individuals into the different latent types.

For this study we use nationally representative data from the UKHLS. We combine a rich set of nurse-measured and non-fasted blood-based biomarkers to build a cumulative risk score index (also known as allostatic load) which takes into account the chronic exposure to psychosocial and environmental challenges. This allows us to assess the lasting contribution of circumstances and efforts to inequality in long-term health measures.

Our results show a clear ordering of types with respect to both our composite biomarker measure (allostatic load) and the underlying observed circumstances. Beyond the definition of types, FMM analysis allows us to explore the type-specific unobserved heterogeneity in the association between our health measure and efforts, which is crucial for the measurement of *ex post* IOp. Taking advantage of the later along with our latent class approach to define types, we further contribute to the literature by extending a recently developed decomposition technique on IOp in health (Carrieri and Jones, 2018). Our more parsimonious and data-driven definition of types (using a latent class model framework) are of importance given the recent evidence that a large number of types may create upward bias in the IOp measurement (Brunori et al., 2019).

We find that a latent class model with three unobserved types provides the best fit with our data, indicating that a relatively small number of types are enough to characterise the sample. Our results show that the characteristics of each of these types reflect a complex combination of observed circumstances, which may be missed if single circumstances or *ad hoc* selections of circumstances were chosen to define types. After classifying individuals into the latent types using modal assignments, we decompose overall inequality in allostatic load. We find that the sum of all sources of inequality in allostatic load attributable to these types (direct effect of circumstances and indirect via their influence on efforts) is about 63% (about 50% due to direct role of circumstances). On the other hand, legitimate sources of inequality (the direct contribution of efforts), which are consistent with the reward principle, account for only around 3% of the total inequality.

References

Aaberge, R., Mogstad, M., Peragine, V. (2011). Measuring long-term inequality of opportunity. *Journal of Public Economics*, 95, 193-204.

Bago d'Uva, T. and Jones, A.M. (2009) "Health care utilisation in Europe: new evidence from the ECHP", *Journal of Health Economics* 28, 265-279.

Bago d'Uva, T., Jones, A.M. and van Doorslaer, E. (2009) "Measurement of horizontal inequity in health care utilisation using European panel data", *Journal of Health Economics*, 28, 280-289.

Bago d'Uva, T., O'Donnell, O., Van Doorslaer, E. (2008). Differential health reporting by education level and its impact on the measurement of health inequalities among older Europeans. *International Journal of Epidemiology*, 37, 1375-1383.

Balia, S., Jones, A.M. (2011). Catching the habit: a study of inequality of opportunity in smoking-related mortality. *Journal of the Royal Statistical Society Series A*, *174*, 175-194.

Bond, T.N., Lang, K. (2019). The sad truth about happiness scales. *Journal of Political Economy*, forthcoming.

Bourguignon, F., Ferreira, F.H.G., Menendez, M. (2007). Inequality of opportunity in Brazil. *Review of Income and Wealth,* 53, 585-618.

Brunori, P., Peragine, V., Serlenga, L. (2019). Upward and downward bias when measuring inequality of opportunity. *Social Choice and Welfare*, 52: 635-661.

Cameron, A.C. and Trivedi, P.K., 2010. Microeconometrics Using Stata. College Station, TX: Stata press.

Carrieri, V., Jones, A.M. (2018). Inequality of opportunity in health: A decomposition-based approach. *Health Economics*, 27, 1981-1995.

Crossley, T. F., Kennedy, S. (2002). The reliability of self-assessed health status. *Journal of Health Economics*, 21, 643-658.

Davillas, A., Jones, A.M. (2018). Ex ante Inequality of Opportunity in Health, Decomposition and Distributional Analysis of Biomarkers (No. 18/30). HEDG, c/o Department of Economics, University of York.

Davillas, A., Pudney, S. (2017). Concordance of health states in couples: analysis of self-reported, nurse administered and blood-based biomarker data in the UK understanding society panel. *Journal of Health Economics*, 56, 87-102.

Deb, P., Gallo, W.T., Ayyagari, P., Fletcher, J.M., and Sindelar, J.L. (2011). The effect of job loss on overweight and drinking. *Journal of Health Economics*, 30, 317-327.

Deutsch, J., Alperin, M.N.P., Silber, J. (2018). Using the Shapley Decomposition to Disentangle the Impact of Circumstances and Efforts on Health Inequality. *Social Indicators Research*, 138, 523-543.

Ferreira, F.H., Gignoux, J. (2013). The measurement of educational inequality: Achievement and opportunity. *The World Bank Economic Review*, 28, 210-246.

Fleurbaey, M., Peragine, V. (2013). Ex ante versus ex post equality of opportunity. *Economica*, 80, 118-130.

Fleurbaey, M., Schokkaert, E. (2009). Unfair inequalities in health and health care. *Journal of Health Economics*, 28, 73-90.

Fleurbaey, M., Schokkaert, E. (2012). Equity in health and health care, in Barros, P., McGuire T., Pauly, M. (eds.), *Handbook of Health Economics, Volume 2*, 1003-1092.

Garcia-Gomez, P., Schokkaert, E., van Ourti, T., Bago d'Uva, T. (2015). Inequity in the face of death. *Health Economics,* 24, 1348-1367.

Howard, J. T., and Sparks, P. J. (2016). The effects of allostatic load on racial/ethnic mortality differences in the United States. *Population Research and Policy Review*, 35(4), 421-443.

Jusot, F., Tubeuf, S., Trannoy, A. (2013). Circumstances and efforts: how important is their correlation for the measurement of inequality of opportunity in health? *Health Economics*, 22, 1470-1495.

Kenkel, D., Lillard, D., Mathios, A. (2006). The roles of high school completion and GED receipt in smoking and obesity. *Journal of Labor Economics*, 24, 635-660.

Lefranc, A., Pistolesi, N., Trannoy, A. (2009). Equality of opportunity and luck: definitions and testable conditions, with an application to income in France. *Journal of Public Economics*, 93, 1189-1207.

Lefranc, A., Trannoy, A. (2017). Equality of opportunity, moral hazard and the timing of luck. *Social Choice and Welfare*, 49(3-4), 469-497.

Li Donni, P., Rodriguez, J.G., Rosa Dias, P., (2015). Empirical definition of social types in the analysis of inequality of opportunity: a latent class approach. *Social Choice and Welfare*, 44, 673-701.

McEwen, B.S., (2015). Biomarkers for assessing population and individual health and disease related to stress and adaptation. *Metabolism*, 64(3), S2-S10.

NHS England (2017). NHS England's response to the specific equality duties of the Equality Act 2010. London: NHS England.

Ramos, X., Van de Gaer, D. (2016) Approaches to inequality of opportunity: principles, measures and evidence. *Journal of Economic Surveys*, 30, 855-883.

Roemer, J.E. (1998). *Equality of opportunity*. Harvard University Press.

Roemer, J.E. (2002). Equality of opportunity: A progress report. *Social Choice and Welfare,* 19, 455-471.

Roemer, J.E., Trannoy, A. (2016). Equality of opportunity: theory and measurement, *Journal of Economic Literature,* 54, 1288-1332.

Rosa Dias, P. (2009). Inequality of opportunity in health: evidence from a UK cohort study. *Health Economics*, 18, 1057-1074.

Rosa Dias, P. (2010). Modelling opportunity in health under partial observability of circumstances. *Health Economics*, 19, 252-264.

Seeman, T.E., Crimmins, E., Huang, M. H., Singer, B., Bucur, A., Gruenewald, T., Berkman L.F., Reuben, D.B. (2004). Cumulative biological risk and socio-economic differences in mortality: MacArthur studies of successful aging. *Social Science & Medicine*, 58, 1985-1997.

Shorrocks, A.F. (1982). Inequality decomposition by factor components. *Econometrica*, 193-211.

Trannoy, A., Tubeuf, S., Jusot, F., Devaux, M. (2010). Inequality of opportunities in health in France: a first pass. *Health Economics*, 19, 921-938.

Vermunt, J.K., Magidson, J. (2004). Latent class analysis. *The Sage Encyclopedia of Social Sciences Research Methods*, 2, 549-553.

Appendix

Table A1. Description of biomarkers used for allostatic load

Biomarker	Description	Mean	Standard deviation
Waist-to-height ratio (WHR)	Waist circumference (cm) over height (cm)	0.560	0.079
Systolic blood pressure (SBP)	Maximum pressure in an artery when the heart is pumping blood (mmHg)	126.4	16.3
C-reactive protein (CRP)	Inflammatory biomarker; rises as part of the immune response to infection (mg/L)	2.028	1.957
Fibrinogen	Fibrinogen (g/L) is a glycoprotein that aids the body to stop bleeding by promoting blood clotting, and is regarded as an inflammatory biomarker.	2.744	0.511
Glycated haemoglobin (HbA1c)	Blood sugar biomarker; diagnostic test for diabetes. (mmol/mol)	36.8	6.4
Cholesterol ratio	Fat in the blood biomarker; ratio of the total cholesterol (mmol/L) over the high-density lipoprotein cholesterol (mmol/L).	3.717	1.304

Table A.2. Posterior probabilities and mean allostatic load by latent class (type): FMMs with different number of latent classes.

Number of latent classes (types)		Posterior type membership probabilities (%)	Predicted mean of allostatic load
K=2	Type 1	29.93 (27.46; 32.52)	24.34 (24.14; 24.54)
	Type 2	70.07 (67.48;72.54)	28.50 (28.39; 28.61)
K=3	Type 1	20.63 (17.10; 24.67)	23.50 (23.16; 23.84)
	Type 2	40.35 (33.20; 47.94)	26.63 (26.23; 27.04)
	Type 3	39.02 (29.61; 49.32)	29.89 (29.35; 30.43)
K=4	Type 1	15.77 (11.45; 21.34)	23.03 (22.57; 23.50)
	Type 2	28.43 (21.57; 36.47)	25.70 (25.16; 26.23)
	Type 3	45.14 (34.71; 56.02)	28.45 (27.76; 29.13)
	Type 4	10.64 (5.19; 20.58)	32.68 (31.49; 33.87)
K=5	Type 1	19.43 (15.09; 24.65)	23.41 (23.00; 23.81)
	Type 2	29.06 (20.01; 40.15)	26.19 (25.71; 26.67)
	Type 3	15.15 (5.57; 35.10)	27.35 (26.68; 28.01)
	Type 4	1.16 (0.69; 1.93)	28.13 (28.07; 28.18)
	Type 5	35.20 (24.20; 48.03)	30.13 (29.33; 30.92)
K=6	Type 1	7.64 (3.66; 15.28)	22.37 (21.73; 23.01)
	Type 2	19.71 (14.76; 25.82)	24.73 (23.96; 25.49)
	Type 3	13.79 (3.71; 39.91)	26.93 (24.23; 29.63)
	Type 4	35.64 (13.57; 66.14)	27.37 (26.15; 28.60)
	Type 5	1.56 (0.64; 3.79)	30.79 (28.97; 32.61)
	Type 6	21.64 (6.81; 51.07)	31.04 (28.50; 33.59)
K=7	Type 1	14.69 (8.30; 24.65)	22.94 (22.18; 23.71)
	Type 2	22.84 (17.92; 28.64)	25.49 (24.61; 26.38)
	Type 3	0.97 (0.75; 1.27)	25.86 (25.80; 25.91)
	Type 4	2.46 (1.01; 5.83)	27.24 (25.92; 28.56)
	Type 5	14.70 (8.24; 24.81)	28.17 (27.37; 28.96)
	Type 6	32.97 (24.03; 43.34)	28.23 (27.16; 29.29)
	Type 7	11.37 (4.22; 27.20)	32.54 (30.80; 34.28)

Notes: 95% Confidence intervals in parenthesis

Table A3. Heterogeneous association between efforts and allostatic load by latent type: FMM estimates with three latent types.

	Type 1	Type 2	Type 3
Current smoker	0.369**	0.826***	1.163***
	(0.161)	(0.183)	(0.204)
Ex-smoker	-0.005	0.332***	0.394***
	(0.121)	(0.110)	(0.142)
Non-compliance:5 fruits/vegetables/day	-0.125	0.182	0.259*
	(0.134)	(0.120)	(0.154)
White bread	0.158	0.305**	0.540***
	(0.133)	(0.125)	(0.150)
Non-frequent walking	-0.014	0.039	0.623***
	(0.113)	(0.113)	(0.138)
Sports activity: 1-3 times/week	-0.045	0.519***	0.133
	(0.164)	(0.173)	(0.240)
Sports activity: at least once/month	0.026	1.033***	0.509*
	(0.178)	(0.206)	(0.281)
Sports activity: less frequent/not at all	0.257*	1.104***	0.974***
	(0.155)	(0.153)	(0.211)
Constant term	23.380***	25.341***	28.311***
	(0.186)	(0.255)	(0.338)

* p-value<0.1; ** p-value<0.05; *** p-value<0.01.