Annals of Actuarial Science (2025), **19**, pp. 416–441
doi:10.1017/S1748499524000320

ORIGINAL RESEARCH PAPER

A compositional approach to modeling cause-specific mortality with zero counts

Zhe Michelle Dong[1], Han Lin Shang[2], Francis Hui[1] and Aaron Bruhn[1]

[1]Research School of Finance, Actuarial Studies and Statistics, The Australian National University, Canberra, Australia; and [2]Department of Actuarial Studies and Business Analytics, Macquarie University, Sydney, Australia
Corresponding author: Zhe Michelle Dong; Email: zhe.dong@anuedu.au

(Received 30 January 2024; revised 06 December 2024; accepted 18 December 2024; first published online 17 January 2025)

Abstract

Understanding and forecasting mortality by cause is an essential branch of actuarial science, with wide-ranging implications for decision-makers in public policy and industry. To accurately capture trends in cause-specific mortality, it is critical to consider dependencies between causes of death and produce forecasts by age and cause coherent with aggregate mortality forecasts. One way to achieve these aims is to model cause-specific deaths using compositional data analysis (CODA), treating the density of deaths by age and cause as a set of dependent, nonnegative values that sum to one. A major drawback of standard CODA methods is the challenge of zero values, which frequently occur in cause-of-death mortality modeling. Thus, we propose using a compositional power transformation, the α-transformation, to model cause-specific life-table death counts. The α-transformation offers a statistically rigorous approach to handling zero value subgroups in CODA compared to *ad hoc* techniques: adding an arbitrarily small amount. We illustrate the α-transformation in England and Wales and US death counts by cause from the Human Cause-of-Death database, for cardiovascular-related causes of death. The results demonstrate the α-transformation improves forecast accuracy of cause-specific life-table death counts compared with log-ratio-based CODA transformations. The forecasts suggest declines in the proportions of deaths from major cardiovascular causes (myocardial infarction and other ischemic heart diseases).

Keywords: Compositional data analysis; cause of death; log-ratio transformation; alpha transformation; mortality forecasting

1. Introduction

Understanding mortality by cause is key to informing medical research decisions and planning social services (Alai et al., 2015; Kjaergaard et al., 2019). It is also important in assessing mortality rates and longevity risk for life insurers, as causal factors can drive the best estimate of mortality and morbidity assumptions for the purposes of reserving and pricing. Analyzing and modeling cause-of-death data present two main challenges: the need to account for inherent dependencies between various causes of deaths and the need to produce forecasts by causes coherent with aggregate mortality forecasts.

Traditional methods for mortality modeling and forecasting, including the Lee-Carter (LC) model (Lee & Carter, 1992) or variations thereof and the age-period-cohort model (Holford, 1983; Renshaw & Haberman, 2006), among others, generally do not account for dependencies between competing causes of death. As such, over the past two decades, considerable progress has been made on joint models for multiple causes of death, which capture between-cause dependencies. Arnold and Sherris (2013) applied vector error correction models to cause-of-death

© The Author(s), 2025. Published by Cambridge University Press on behalf of Institute and Faculty of Actuaries. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S1748499524000320 Published online by Cambridge University Press

mortality rates to quantify the dependence between competing risks and subsequently found an improvement in forecasts compared to methods that do not allow for such dependencies. Alai *et al.* (2015) formulated a multinomial logistic model across several causes of death to investigate the effects of improvement and elimination of mortality due to cancer. Li *et al.* (2019) adopted a forecast reconciliation approach to ensure coherence in cause-specific mortality rates, while Li and Lu (2019) introduced hierarchical Archimedean copulas to capture dependence between competing risks in causes of death. More recently, Zhang *et al.* (2023) developed a predictive approach for cause-of-death mortality modeling that jointly models various causes, ages, and years using a penalized tensor decomposition.

The majority of literature on modeling mortality by cause, including those mentioned above, treats cause-specific life-table deaths as non-compositional, that is, through modeling age-specific mortality rates rather than age distributions of death. Although these methods enable modeling dependencies between mortality rates for different causes, a more direct approach is to forecast the cause-specific death distribution, where the dependence is explicitly incorporated by capturing relativities between deaths of one cause and another. Indeed, cause-of-death data are fundamentally compositional, as deaths have been recorded and attributed to various causes for analysis in globally used medical classifications for epidemiology, health management, and understanding mortality experience (World Health Organization, 1992).

With this in mind, an alternative approach, known as compositional data analysis (CODA), has arisen in the actuarial science literature, which aims to model the cause-specific death distribution directly and produce mortality forecasts arising from the composition of the distribution itself. The idea of CODA dates back to the seminal work of Aitchison (1982) for analyzing data that arise as a vector of observations where the elements sum to a constant value and, therefore, only contain relative information. In the context of mortality by cause, the compositional sum constraint translates to avoided deaths from one cause, leading to increased deaths from other causes.

When analyzing compositional data, methods that ignore the compositional constraint and apply standard multivariate data analysis to the raw observations (we refer to such approaches as "raw data analysis" or RDA) can encounter potential issues with coherence when it comes to aggregated mortality forecasts. An alternative approach in CODA is to transform the compositional data from the simplex, subject to the unit sum constraint, to the unconstrained real space before applying standard multivariate data analysis and forecasting. Then, the results are transformed into the compositional space for interpretation and inference. Within this latter approach, log-ratio transformations are by far the most widely used to transform compositional data due to their various attractive compositional properties (see Aitchison, 1982, for details). The first to propose such a "log-ratio analysis" or LRA for forecasting mortality rates was Oeppen (2008), who applied an LC mortality model to log-ratio transformed death compositions to forecast cause-specific mortality. Oeppen (2008) used centered log-ratio (CLR) transformation (see Section 2 for details) and found that capturing dependencies between subgroups via LRA and the CODA framework improved the overall forecast while assuming independence between causes tended to produce pessimistic results; that is, expected deaths tend to be overstated. Kjaergaard *et al.* (2019) further extended this approach by developing two new LRA models for cause-specific deaths, adding cause-specific weights to age and time subgroups, and decomposing joint and individual variation between causes of death to improve forecast accuracy further. Other notable works include that of Bergeron-Boucher *et al.* (2017), who applied CODA to produce age-coherent forecasts for mortality; Bergeron-Boucher *et al.* (2022), who used LRA to model healthy life expectancy; and Kjaergaard *et al.* (2020), who produced longevity forecasts by socioeconomic group using LRA.

While the aforementioned works use LRA to address some of the issues with analyzing compositional data (relative to RDA), one outstanding challenge with LRA-based modeling is the presence of zero counts/values (Bergeron-Boucher et al., 2017; Kjaergaard et al., 2019, 2020). Specifically, compositional data with zero values can be interpreted as lying on a boundary of

https://doi.org/10.1017/S1748499524000320 Published online by Cambridge University Press

the simplex. So, naively applying a log-ratio transformation to such data results in one or more transformed values taking $\pm\infty$. In the context of mortality by cause, zero death counts in subcategories of the composition arise commonly for new and emerging or granular causes of death at certain ages and at older ages where exposure is limited. Since the existence and treatment of zeros may lead to differences in the overall inference and forecasts, as mentioned above, this could have consequences on our understanding of longevity risk and mortality improvements, along with associated financial implications (Basel Committee on Banking Supervision, 2013).

In the literature, the problem of zeros when using LRA has often been addressed in an *ad hoc* manner by omitting, aggregating, or adding small arbitrary values to zero values (Martin-Fernandez et al., 2003). For instance, Kjaergaard *et al.* (2019) explored imputing half of the minimum observed death count, a method initially used by Bergeron-Boucher *et al.* (2017). Alternatively, Kjaergaard *et al.* (2019) noted that Hyndman *et al.* (2013) imputed death rates based on information from nearby years for the same age group using linear interpolation. None of these methods is ideal; furthermore, Greenacre (2021) compared four different algorithms to substitute zeros and showed the resulting conclusions could be susceptible to the technique of zero substitution. More recently, Greenacre (2024) introduced the χ-power transformation to address the problem of zeros in compositional data by combining the chi-squared distance in correspondence analysis with the Box-Cox power transformation.

In this article, we propose a novel approach to modeling mortality by cause with zero values using a modification of LRA. We introduce a compositional power transformation known as the α-transformation (Tsagris et al., 2011), which addresses the challenges presented by zero values in the setting of CODA in a more statistically principled manner compared to the aforementioned *ad hoc* techniques. The α-transformation, which maps compositional data to remove their unit sum constraint, is a generalized Box-Cox power transformation that includes both RDA and LRA as special cases but more broadly involves a tuning parameter $\alpha \in (0, 1]$. This parameter can be calibrated in a data-driven manner to enable more flexibility in producing forecasts compared to standard LRA when there are zero values in the data. While the α-transformation has been applied to CODA for geology and biology, among other fields (Tsagris & Stewart, 2020), to our knowledge, this paper is the first to examine its use in forecasting mortality by age and cause.

We apply the α-transformation to two datasets: 16 years of cause-of-death data from England and Wales data and 43 years of cause-of-death data from the USA. In both applications, we disaggregate for cardiovascular causes such that there are data with zero counts in one or more subgroups. We couple the α-transformation with the LC mortality model for multivariate analysis and forecasting (similar to those of Kjaergaard et al., 2019; Oeppen, 2008), and compare results with several LRA and RDA approaches where *ad hoc* methods are used to deal with zero values. The results across both applications demonstrate that the α-transformation generally improves mortality forecast by cause, while having the added benefit of being able to analyze compositional data with zero counts in a rigorous yet data-driven manner. The α-transformation is shown to address the key issue of zero counts in mortality data, generalizing the log-ratio transformation to a broader class of transformations and providing additional flexibility and improved performance when forecasting mortality by cause using CODA-based techniques.

The remainder of this paper is structured as follows: Section 2 reviews several key ideas, including the Lee-Carter (LC) mortality model and LRA. Section 3 introduces the α-transformation for mortality by cause data. Section 4 applies the proposed methodology to forecast mortality on cause-of-death data from England and Wales and the USA, while Section 5 offers some concluding remarks.

2. Review of key concepts

We review three foundational concepts for understanding how the α-transformation can be applied to cause-of-death mortality modeling, namely, compositional data (Section 2.1), log-ratio analysis or LRA (Section 2.2), and the LC mortality model (Section 2.3).

https://doi.org/10.1017/S1748499524000320 Published online by Cambridge University Press

2.1 Compositional data

Cause-specific mortality can be represented by actual death counts per combination of year, age group, and cause. Specifically, let $D_{t,u,c}$ denote the actual death count for year $t = 1, 2, \ldots, T$, age group $u = 1, 2, \ldots, U$, and cause $c = 1, 2, \ldots, C$, and define $D_t = \sum_{u=1}^{U} \sum_{c=1}^{C} D_{t,u,c}$ as the total deaths across all age bands and cause groups for year t. Then we can calculate $d_{t,u,c} = D_{t,u,c}/D_t$ such that for a given year, the vector $\boldsymbol{d}_t = (d_{t,1,1}, d_{t,1,2}, \ldots, d_{t,1,C}, d_{t,2,1}, d_{t,2,2}, \ldots, d_{t,2,C}, d_{t,u,1}, d_{t,u,2}, \ldots, d_{t,U,C})$ represents the density distribution of deaths by age group and cause. The densities in $\boldsymbol{d}_t$ are ordered such that the cause runs faster than age. Moreover, the compositional vector satisfies $\sum_{u=1}^{U} \sum_{c=1}^{C} d_{t,u,c} = 1$. Moreover, by stacking the $\boldsymbol{d}_t$'s as row vectors on top of each other, we can form the $T \times UC$ compositional matrix $\mathbf{D}$ of death densities

$$\mathbf{D} = \begin{pmatrix} d_{1,1,1} & d_{1,1,2} & \ldots & d_{1,1,C} & d_{1,2,1} & d_{1,2,2} & \ldots & d_{1,U,C} \\ d_{2,1,1} & d_{2,1,2} & \ldots & d_{2,1,C} & d_{2,2,1} & d_{2,2,2} & \ldots & d_{2,U,C} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{T,1,1} & d_{T,1,2} & \ldots & d_{T,1,C} & d_{T,2,1} & d_{T,2,2} & \ldots & d_{T,U,C} \end{pmatrix}. \tag{1}$$

Due to the sum-to-one constraint, only $UC - 1$ elements are needed to uniquely determine each vector $\boldsymbol{d}_t$. Statistically then, the sample space for compositional cause-of-death mortality data is a simplex: for all $t = 1, \ldots T$,

$$S^{UC-1} = \left\{ (d_{t,1,1}, \ldots, d_{t,U,C}) | d_{t,u,c} \geq 0, \quad \sum_{u=1}^{U} \sum_{c=1}^{C} d_{t,u,c} = 1 \right\}.$$

2.2 Log-ratio analysis

A common approach to analyzing compositional data is to employ the log-ratio transformations class, which seeks to transform the data from the simplex back to an unconstrained real space before building a statistical model for analysis. The two most common types of transformations within LRA are the CLR and isometric log-ratio (ILR) transformations, which we consider in this paper. Importantly, the CLR and ILR are used for analyzing compositional data *without* zero values.

The CLR transformation is defined by dividing all the values in the compositional vector by their geometric mean before applying the natural log transformation. For row t in (1), the CLR for each element is given by

$$w(d_{t,u,c}) = \ln \left(\frac{d_{t,u,c}}{(\prod_{u=1}^{U} \prod_{c=1}^{C} d_{t,u,c})^{1/UC}} \right) = \ln (d_{t,u,c}) - \frac{1}{UC} \sum_{u=1}^{U} \sum_{c=1}^{C} \ln (d_{t,u,c}). \tag{2}$$

The CLR transformation is symmetric relative to the compositional parts and has the same number of components as the number of parts in the original composition. We can express the CLR-transformed vector as $\boldsymbol{w}(\boldsymbol{d}_t) = (w(d_{t,1,1}), w(d_{t,1,2}), \ldots, w(d_{t,U,C}))$, noting distances between any two elements of this vector remain the same when measured in the simplex and the real space, thus making the CLR particularly useful for analysis (Grifoll et al., 2019). While each element is no longer constrained to be nonnegative (in principle, they can take any real number), the entire vector remains constrained since the elements must sum to zero by the construction of (2).

To further remove this constraint, the ILR left matrix multiplies the CLR-transformed vector by a Helmert sub-matrix and has been promoted as the more theoretically correct method (especially to contrast groups of elements) in CODA (Greenacre & Grunsky, 2019). The Helmert sub-matrix is an orthonormal $(UC - 1) \times UC$ matrix formed by deleting the first row of the Helmert orthogonal matrix (see Greenacre, 2021, and Tsagris & Stewart, 2022, for technical details). If we denote

https://doi.org/10.1017/S1748499524000320 Published online by Cambridge University Press

this Helmert sub-matrix as $\boldsymbol{H}$, then the ILR-transformed vector is defined as

$$\boldsymbol{z}(\boldsymbol{d}_t) = \boldsymbol{H}\boldsymbol{w}(\boldsymbol{d}_t), \tag{3}$$

and is no longer subject to any constraint. That is $\boldsymbol{z}(\boldsymbol{d}_t) \in \mathcal{R}^{UC-1}$, and all of its elements can take any real value.

The CLR and ILR aim to transform compositional data into real unconstrained space. On the other hand, as both these transformations are based on taking logarithms, then such methods will not work if one or more of the actual death counts, and subsequently one or more of the $d_{t,u,c}$'s, are exactly zero in value. This is the motivating problem for our subsequent developments as, in practice, many datasets of death counts tend to include zeros for some cause and age combinations.

2.3 The Lee-Carter model for compositional data

We describe a modification of the LC model introduced by Oeppen (2008) for compositional data. We refer to this model as the LC-CODA model, and its construction can be summarized in the following steps.

(I) Center each row of $\boldsymbol{D}$ in (1) by taking the inverse perturbation of the geometric mean from each row of death densities. This results in a matrix of centered death densities, denoted here as $\widetilde{\boldsymbol{D}}$.
(II) Apply the CLR transformation to each row of $\widetilde{\boldsymbol{d}}$, mapping the vector of UC-compositions for a given year t from the simplex to a UC-dimensional Euclidean subspace.
(III) Fit and forecast the transformed data using the LC model. Note other more sophisticated models are possible here (e.g., Bergeron-Boucher et al., 2017; Kjaergaard et al., 2019, 2020), and this step and all our developments can be modified to employ such approaches. For simplicity, though, we focus on the LC model.
(IV) Back-transform the estimated death densities to the simplex by inverting the CLR transformation and performing a compositional perturbation to the geometric mean for each row estimate to obtain the final forecasted compositional results.

We elaborate each of the steps above in detail. Consider the matrix of compositional death densities in (1), and compute $\mathbf{g}$ as the UC-vector, the elements of which are given by the column-wise geometric mean of $\mathbf{D}$, that is, $\mathbf{g} = ((\prod_{t=1}^{T} d_{t,1,1})^{1/T}, (\prod_{t=1}^{T} d_{t,1,2})^{1/T}, \ldots, (\prod_{t=1}^{T} d_{t,U,C})^{1/T})$. Next, define the perturbation operation and its inverse as follows (Aitchison, 1982). For two vectors of compositions $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_n)$, all of the elements of which are nonzero, we have

$$\text{Perturbation:} \quad X \oplus Y = C\left(x_1 y_1, x_2 y_2, \ldots, x_n y_n\right)$$
$$\text{Inverse perturbation:} \quad X \ominus Y = C\left(\frac{x_1}{y_1}, \frac{x_2}{y_2}, \ldots, \frac{x_n}{y_n}\right),$$

where the operator $C(\cdot)$ "closes" the row, that is, normalizes by dividing each entry by the sum of all entries.

In Step (I) of fitting the LC-CODA model, we apply a centering process to construct a matrix of centered death densities, $\widetilde{\boldsymbol{D}}$, where the tth row of $\widetilde{\boldsymbol{D}}$ for $t = 1, \ldots, T$ is computed as

$$\widetilde{\boldsymbol{d}}_t = \boldsymbol{d}_t \ominus \mathbf{g} = C\left(\frac{d_{t,1,1}}{(\prod_{t=1}^{T} d_{t,1,1})^{1/T}}, \frac{d_{t,1,2}}{(\prod_{t=1}^{T} d_{t,1,2})^{1/T}}, \ldots, \frac{d_{t,U,C}}{(\prod_{t=1}^{T} d_{t,U,C})^{1/T}}\right). \tag{4}$$

Note the elements in $\mathbf{g}$ can be considered analogues of the age- and cause-specific average mortality over time in a standard LC model.

https://doi.org/10.1017/S1748499524000320 Published online by Cambridge University Press

In Step (II), we apply the CLR transformation to obtain the vector $\boldsymbol{w}(\widetilde{\boldsymbol{d}}_t) = (w(\widetilde{d}_{t,1,1}), w(\widetilde{d}_{t,1,2}), \ldots, w(\widetilde{d}_{t,U,C}))$ for $t = 1, \ldots, T$, where to be clear the elements are computed using (2) except replacing $d_{t,u,c}$ with $\widetilde{d}_{t,u,c} = d_{t,u,c}/(\prod_{t=1}^{T} d_{t,u,c})^{1/T}$. Let $\boldsymbol{w}(\widetilde{\boldsymbol{D}})$ denote the resulting $T \times UC$ matrix formed by stacking the $\boldsymbol{w}(\widetilde{\boldsymbol{d}}_t)$'s as row vectors on top of one another.

In Step (III), we apply the singular value decomposition to $\boldsymbol{w}(\widetilde{\boldsymbol{D}})$ and estimate the LC mortality model analogous to how it is done for the non-compositional setting. We provide details of this in Appendix A.1, but to summarize, we fit a model of the form

$$w(\widetilde{d}_{t,u,c}) = b_{u,c}k_{t,c} + \epsilon_{t,u,c}, \tag{5}$$

where $b_{u,c}$ denotes age- and cause-specific coefficients that vary over time, $k_{t,c}$ denotes factors of time-varying indices for the level of mortality, and $\epsilon_{t,u,c}$ denotes a residual error term. Note that a mean/intercept term is omitted from (5) due to centering from the geometric mean in Step (I). For forecasting, we can adopt a similar approach to Kjaergaard *et al.* (2019) and Zhang *et al.* (2023), among others, who applied time-series methods such as random walk with drift to $k_{t,c}$, and substitute forecasted values of these back in to (5).

Finally, in Step (IV), after obtaining forecasted values of $w(\widetilde{d}_{t,u,c})$, we can apply an inverse CLR transformation followed by a perturbation operation to obtain the actual forecasted death density distribution. In detail, suppose that at a future time $T' > T$, the predicted value of the time factor for cause c is given by $\widehat{k}_{T',c}$, while the estimated age- and cause-specific coefficients from Step (III) are given by $\widehat{b}_{u,c}$. Then a vector of forecasted centered death densities is given by $\widetilde{\boldsymbol{d}}_{T'} = (\widetilde{d}_{T',1,1}, \widetilde{d}_{T',1,2}, \ldots, \widetilde{d}_{T',U,C})$ where $\widetilde{d}_{T',u,c} = w^{-1}(\widehat{b}_{u,c}\widehat{k}_{T',c})$ and $w^{-1}(\cdot)$ denotes the inverse CLR transformation. The corresponding vector for forecasted death densities from the LC-CODA model is then given by $\widehat{\boldsymbol{d}}_{T'} = \widetilde{\boldsymbol{d}}_{T'} \oplus \boldsymbol{g}$.

Compared to modeling mortality rates independently, one key element of using the LC-CODA model is that death counts are naturally redistributed through compositional constraints. As mortality changes over time, if some deaths do not occur at a specific age band and cause, they are naturally shifted toward a different age band and cause group. This maintains subcompostional coherence with the total number of deaths per year as given by the initial life table and ensures the disaggregated death forecasts will be coherent with the overall aggregated mortality forecast (Oeppen, 2008). In the context of compositional data, subcompositional coherence refers to the property that relationships between parts of a composition are unaffected by forming subcompositions, such that results and summary statistics based on the subcomposition are the same as the composition (Greenacre, 2021). On the other hand, due to its reliance on the CLR transformation, the LC-CODA model is unable to handle zero values in the raw densities $d_{t,u,c}$, and these would need to be omitted, aggregated, or replaced with an arbitrarily small value before step (I).

We present a more detailed exposition of LC-CODA in Appendix A.1, which we use in the application of the α-transformation and log-ratio transformations in this paper.

3. Mortality by cause using the α-transformation

Motivated by the challenges of applying LRA to cause-of-death mortality modeling where there are one or more zero values in the death densities, we propose using the α-transformation before applying the LC-CODA model for forecasting.

The α-transformation can be viewed as a Box-Cox transformation applied to the ratios of components, where $\alpha \in (0, 1]$ is a tuning parameter that is tuned to handle compositional challenges in the data with zeros (Tsagris et al., 2011). In detail, let $w^{\alpha}(x)$ represent the Box-Cox transform of a random variable x (Box & Cox, 1964),

$$w^{\alpha}(x) = \begin{cases} \ln(x) & \alpha = 0 \\ \frac{x^{\alpha}-1}{\alpha} & \alpha \neq 0, \end{cases}$$

https://doi.org/10.1017/S1748499524000320 Published online by Cambridge University Press

and recall the matrix of centered death densities $\widetilde{\boldsymbol{D}}$ in (4). For row $t = 1, \ldots, T$, the α-transformation is then defined as

$$\boldsymbol{z}^{\alpha}(\widetilde{\boldsymbol{d}}_t) = \boldsymbol{H}\boldsymbol{w}^{\alpha}(\widetilde{\boldsymbol{d}}_t), \tag{6}$$

where $\boldsymbol{H}$ is the Helmert sub-matrix defined as part of the ILR transformation in (3) and $\boldsymbol{w}^{\alpha}(\widetilde{\boldsymbol{d}}_t)$ denotes the vector where the Box-Cox transformation is applied to each element of $\widetilde{\boldsymbol{d}}_t$. That is, $w^{\alpha}(\widetilde{d}_{t,u,c}) = \ln(\widetilde{d}_{t,u,c})$ if $\alpha = 0$, otherwise $w^{\alpha}(\widetilde{d}_{t,u,c}) = (UC)(\widetilde{d}^{\alpha}_{t,u,c} - 1)/\alpha$ for $\alpha \neq 0$. Note when $\alpha = 0$, the transformation reduces to the ILR transformation defined in (3). If there is no left matrix multiplication by the Helmert sub-matrix H, then we obtain the CLR in (2). Critically, when α is restricted to be greater than zero, the transformed values are well defined even when the raw death densities $\widetilde{d}_{t,u,c} = 0$. This differs from both the ILR and CLR, neither of which can be computed for zero values.

The corresponding sample space of the α-transformation is known as the α space, which we denote as $\mathbb{A}_{\alpha}^{UC-1}$ and is given by

$$\mathbb{A}_{\alpha}^{UC-1} = \left\{ \boldsymbol{z}^{\alpha}(\widetilde{\boldsymbol{d}}_t) \middle| -\frac{1}{\alpha} \leq w^{\alpha}(\widetilde{d}_{t,u,c}) \leq \frac{(UC-1)}{\alpha}, \quad \sum_{u=1}^{U}\sum_{c=1}^{C} w^{\alpha}(\widetilde{d}_{t,u,c}) = 0 \right\}.$$

It is not difficult to see that, similar to the ILR transformation, the vectors in $\mathbb{A}_{\alpha}^{UC-1}$ are not subject to the zero-sum constraint. As $\alpha \to 0$, then $\mathbb{A}_{\alpha}^{UC-1}$ tends to the $(UC-1)$ dimensional real space $\mathcal{R}^{UC-1}$; this is again consistent with the ILR, except now zero values of death densities can be handled provided $\alpha \neq 0$ (Tsagris & Stewart, 2022). On the other hand, when $\alpha = 1$, the α-transformation is equivalent to RDA, that is, the same as applying standard multivariate analysis ignoring the compositional constraint. While α is often determined using a data-driven approach through maximum likelihood estimation (Tsagris et al., 2011), for strong forecasting performance, in Section 4.1, we discuss an alternative method based on minimizing out-of-sample prediction accuracy.

To construct the LC model in conjunction with the α-transformation, we can apply similar steps to those discussed in Section 2.3, except that Step (II) is modified to Step (IIa) where we apply the α-transformation instead of the CLR, and Step (IV) is modified to Step (IVa) where the transformation back to the simplex requires inverting the α-transformation to obtain the final forecast. With regard to the latter, after forecasting the factors $k_{t,c}$ in a similar manner to Section 2.3, the forecast result derived based on the α-transformed data needs to be mapped back to the compositional simplex.

In detail, at future time $T' > T$ and for $\alpha > 0$, let $\widehat{\boldsymbol{z}}^{\alpha}(\widetilde{\boldsymbol{d}}_{T'}) = (\widehat{z}^{\alpha}(\widetilde{d}_{T',1,1}), \ldots, \widehat{z}^{\alpha}(\widetilde{d}_{T',U,C}))$ denote the vector of forecasted α-transformed centered death densities, where $\widehat{z}^{\alpha}(\widetilde{d}_{T',u,c}) = \widehat{b}_{u,c}\widehat{k}_{T',c}$. Then, the vector of the corresponding inverse α-transformed values is given by $\nu^{\alpha}(\widetilde{\boldsymbol{d}}_{T'}) = \alpha \boldsymbol{H}^{\top}\widehat{\boldsymbol{z}}^{\alpha}(\widetilde{\boldsymbol{d}}_{T'}) + 1$.

Afterward, the forecast vector of death densities at time T' is given by

$$\widetilde{\boldsymbol{d}}_{T'} = \left(\frac{\nu^{1/\alpha}(\widetilde{\boldsymbol{d}}_{t,1,1})}{\sum_{u=1}^{U}\sum_{c=1}^{C} \nu^{1/\alpha}(\widetilde{\boldsymbol{d}}_{T',u,c})}, \ldots, \frac{\nu^{1/\alpha}(\widetilde{\boldsymbol{d}}_{t,U,C})}{\sum_{j=1}^{U}\sum_{k=1}^{C} \nu^{1/\alpha}(\widetilde{\boldsymbol{d}}_{T,u,c})} \right),$$

and $\widehat{\boldsymbol{d}}_{T'} = \widetilde{\boldsymbol{d}}_{T'} \oplus \boldsymbol{g}$.

To conclude, we remark that as long as the forecasted data $\widehat{\boldsymbol{z}}^{\alpha}(\widetilde{\boldsymbol{d}}_{T'})$ lies inside $\mathbb{A}_{\alpha}^{UC-1}$ defined by the original data, then it can be mapped back to the simplex for inference. In some cases during the process of forecasting, for example, for long-term forecasts when $T' \gg T$, it is possible one or more values of $\widehat{\boldsymbol{z}}^{\alpha}(\widetilde{\boldsymbol{d}}_{T'})$ are less than $-1/\alpha$ and lie outside the α-space. This indicates the corresponding forecasts are at or crossing the boundary of the simplex. In such cases, to ensure the

https://doi.org/10.1017/S1748499524000320 Published online by Cambridge University Press

inverse α-transformation is possible, we choose to set corresponding elements of $\widehat{\boldsymbol{z}}^{\alpha}(\widetilde{\boldsymbol{d}}_{T'})$ equal to the boundary value of $-1/\alpha$ (see, e.g., Tsagris et al., 2011, for a similar treatment).

4. Application to the Human Cause-of-Death database

We illustrate an application of the α-transformation coupled with an LC model to cause-of-death counts and life-table deaths for two datasets from England and Wales and the USA as part of the Human Cause-of-Death Data series (HCD, 2024). England and Wales were selected as there have been relatively minimal fluctuations in cause composition during the available data period, while the USA was selected to assess the performance of the proposed α-transformation for a larger dataset spanning more historical years. Disaggregated causes of death within the cardiovascular causes were selected since cardiovascular disease has been steadily decreasing over the past few decades but remains the second-largest cause of death in the UK (British Heart Foundation, 2023; National Institute for Health and Care Excellence, 2023; Raleigh et al., 2022). Data on the complete list of causes of death were obtained, containing 103 causes at the "long" level for England and Wales and 206 causes for US death counts. We treat males and females as separate data sources and perform analysis separately by gender; this is consistent with treatment in earlier CODA literature (e.g., Kjaergaard et al., 2019; Oeppen, 2008).

To perform analysis and forecasting, we aggregated based on age bands and selected causes. We constructed nine age bands: ages 0–24, ages 25–34, ages 35–44, ages 45–54, ages 55–64, ages 65–74, ages 75–84, ages 85–95, and ages over 95. There was an additional age band for the US data comparison, namely, ages 90–99 and then ages over 100. This additional age band was possible due to the availability of the granular death count data from the Human Cause-of-Death Data series (2024) for the USA. Note the age band 0–24 is not a homogeneous group relative to the other age bands, but the reason for aggregating at these ages is twofold: first, for application to life insurance, analysis is typically performed for working age groups; and second, by aggregating across 0–24, there is greater credibility in death counts. We leave the assessment of the variation of deaths by cause at younger ages as an avenue for future investigation.

Turning to causes, for England and Wales' death counts, we aggregated death counts by cause into 11 causes as per the HCD shortlist, with only the cardiovascular causes disaggregated to the "long" list level. For the US death counts, we ensured the same International Classification of Diseases-10 (ICD-10) causes of death were used for comparison. These same cardiovascular causes were mapped to 12 causes as per the HCD "long" list level for the US data. Cardiovascular causes were selected as cardiovascular disease causes of death have steadily decreased over the data period, as introduced at the start of this section. All other causes of death were grouped and aggregated for analysis. The selected cardiovascular causes of death for both datasets are shown in Table 1.

In the disaggregated data for cardiovascular deaths, zero death counts were present across most causes in the disaggregated cardiovascular death category over the available period (2001–2016 for England and Wales and 1979–2021 for the USA,) and when split by age band and across both genders. For example, for England and Wales male data, rheumatic heart disease had zero counts for ages less than 20 (and also for ages 20–30 in 2010) for 2002, 2006, 2010, and 2012–2015. Also, for males, cardiac arrest death counts were zero for ages 40–50 in the year 2004. Similarly, for US male data, acute rheumatic deaths had zero counts for ages less than 20 in 1998, 2002, 2004, 2006–2009, 2014–2016, 2018–2019, and 2021. The same cause had zero counts for ages up to 40 in 2007 and across other older bands in the available years.

In total, for the England and Wales death counts, of the ten cardiovascular causes of death, six had one or more zero counts across both genders in the data: rheumatic heart disease, essential hypertension, hypertensive disease, acute myocardial infarction, cardiac arrest, and heart failure. Not surprisingly, zero death counts for most causes tended to be more prevalent in some years at

https://doi.org/10.1017/S1748499524000320 Published online by Cambridge University Press

Table 1. Selected causes of death, disaggregated for cardiovascular causes, used in our application to England and Wales data (top) and US data (bottom) from the Human Cause-of-Death Data series (2024)

Mortality causes at the "long" level	ICD-10 causes of death
England and Wales data	
48: Rheumatic heart disease	I00–I09
49: Essential hypertension	I10
50: Hypertensive disease (heart, kidney, secondary)	I11–I15
51: Acute myocardial infarction	I21–I23
52: Other IHD	I20, I24, I25
53: Pulmonary heart diseases	I26–I28
54: Non-rheumatic valve disorders	I34–I38
55: Cardiac arrest	I46
56: Heart failure	I50
57: Other heart diseases	I30–I33, I40–I45, I47–I49, I51
1: All other causes of death	All other ICD–10
US data	
102: Acute rheumatic	I00–I02
103: Chronic rheumatic	I05–I09
104: Hypertension	I10
105: Hypertensive (heart)	I11
106: Hypertensive (renal)	I12
107: Hypertensive (both heart and renal)	I13
108: Myocardial Infarction	I21
109: IHD acute	I20, I24
110: IHD chronic	I25
111: Pulmonary	I26–I28
112: Other cardiovascular causes of death	I30–I51
1: All other causes of death	All other ICD–10

younger ages (below 50). Similarly, for US death counts, five of the total 11 cardiovascular causes had zero counts across genders and age bands: acute and chronic rheumatic, hypertension, and hypertensive (both heart and renal). Figures 1 and 2 present aggregated death counts across all ages from 2001 to 2016 for England and Wales and from 1979 to 2021 for US deaths. As observed, the number of deaths for some causes is small, even when aggregated across all ages. With the above in mind, we anticipate forecast performance will improve by explicitly working with actual death counts, that is, including zero values, compared with the standard approach of excluding zeros or replacing them with an arbitrarily small amount.

4.1 Tuning α parameter

To predict cause-of-death data with zero death counts, we proposed selecting an optimal value of α based on out-of-sample forecast accuracy as assessed via an expanding window cross-validation approach. Specifically, for the England and Wales data, as the available data only spanned 16 years, we adopted a simple fourfold expanding window. For the US data, as there were 43 years of data, we adopted a tenfold expanding window. On England and Wales deaths, this meant the first fold consists of the years 2001–2008 for training and 2009–2012 for validation, the second fold consisted of 2001–2009 for training and 2010–2012 for validation (i.e., the training window was

https://doi.org/10.1017/S1748499524000320 Published online by Cambridge University Press

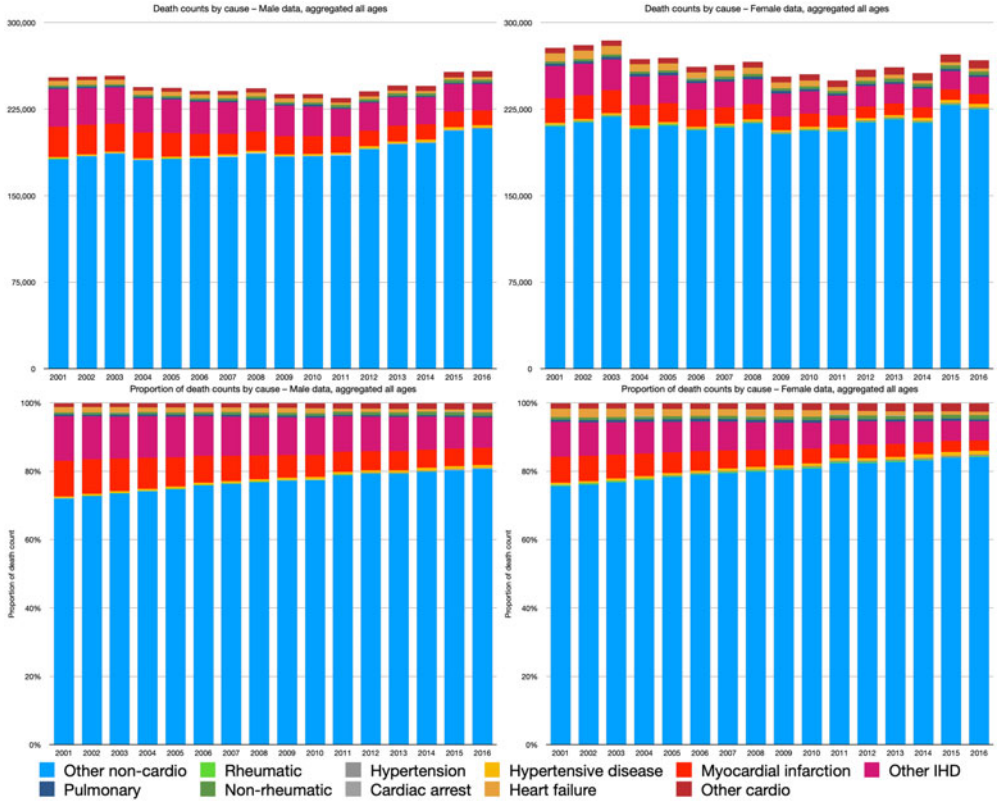


Figure 1. Death counts by cause for England and Wales deaths from 2001 to 2016. The top row presents death counts by cause (disaggregated cardiovascular causes) for males (left) and females (right) in our application to England and Wales data from the Human Cause-of-Death Data series (2024). The bottom row presents the same data but converted to the composition of cardiovascular deaths by cause.

increased by one year) and so on. In each fold, the α-transformation coupled with the LC model as detailed in Section 3 was fitted to the training set, and forecasts were made to the validation set. The years 2013–2016 were held out from all four folds as a test set. Analogously, for the US data, the first fold consisted of the years 1979–2001 for training and 2002–2011 for validation, the second fold consisted of 1979–2002 for training and 2003–2011 for validation, and so on. The years 2012–2021 were held out from all folds as a test set. We remark that as the compositional cause-of-death data exhibits a natural time-series dependence, then an expanding window (or forward chaining) cross-validation method was adopted to tune α; we refer to Racine (2000) and Schnaubelt (2019) for more details around cross-validation in the context of time-series analysis.

For both datasets, we selected α based on minimizing either the average root mean square error (RMSE) or average mean absolute error (MAE) across the four validations sets:

$$\mathrm{RMSE}_k = \sqrt{\frac{\sum_{t=1}^{T_k}\sum_{u=1}^{9}\sum_{c=1}^{11}(\mathrm{observed}_{t,u,c} - \mathrm{predicted}_{t,u,c})^2}{N}};$$

$$\mathrm{MAE}_k = \frac{\sum_{t=1}^{T_k}\sum_{u=1}^{9}\sum_{c=1}^{11}|\mathrm{observed}_{t,u,c} - \mathrm{predicted}_{t,u,c}|}{N},$$

where $\mathrm{observed}_{t,u,c}$ generically denotes the death count for age band u, cause c and the tth year in the validation set, $\mathrm{predicted}_{t,u,c}$ denotes the corresponding predicted death count, and T_k denotes the number of years in the kth validation fold.

https://doi.org/10.1017/S1748499524000320 Published online by Cambridge University Press

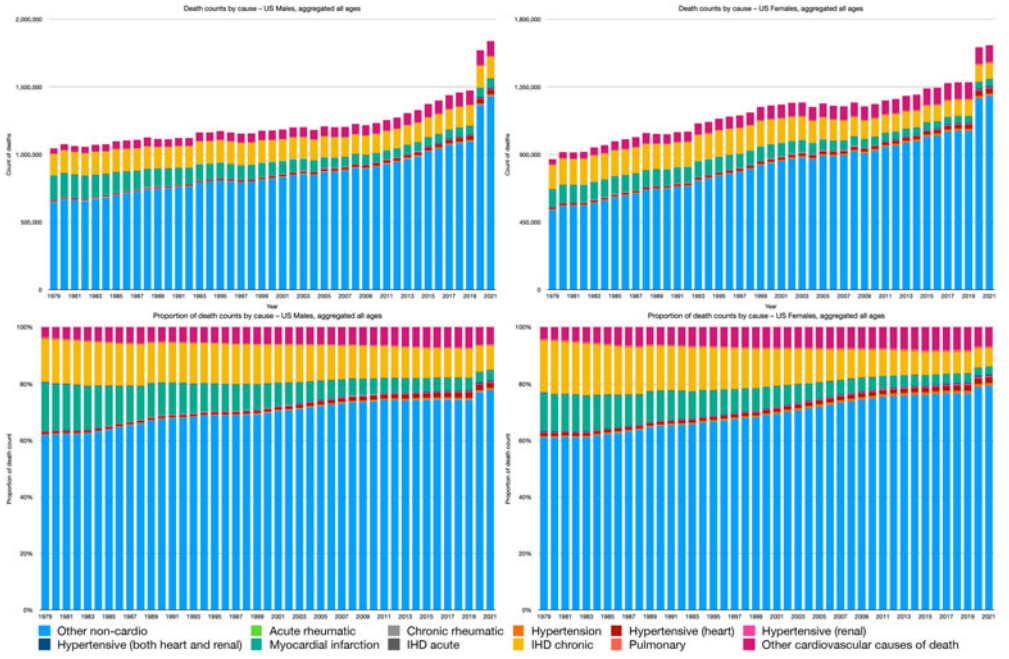


Figure 2. Death counts by cause for US deaths from 1979 to 2021. Note the two years 2020 and 2021 show a spike in deaths, likely due to COVID-19. The top row presents death counts by cause (disaggregated cardiovascular causes) for males (left) and females (right) in our application to England and Wales data from the Human Cause-of-Death Data series (2024). The bottom row presents the same data but converted to the composition of cardiovascular deaths by cause.

Both RMSE and MAE are widely used in model evaluation to measure forecast accuracy (Chai & Draxler, 2014; Hodson, 2022).

$$\text{RMSE} = \frac{1}{4} \times \sum_{k=1}^{4} \text{RMSE}_k$$

$$\text{MAE} = \frac{1}{4} \times \sum_{k=1}^{4} \text{MAE}_k$$

Full results from applying the above cross-validation approach are provided in Appendix A.2. Overall, the optimal α determined using the above cross-validation approach was 0.1 and 0.8 for males and females, respectively, when applied to England and Wales cause-of-death data. On the other hand, optimizing α on the US data yielded values of 0.7 and 0.9, respectively, for males and females. In three of the four cases for optimizing α, the minimum RMSE and MAE produced the same results. Interestingly, the optimal α chosen for the US female data was 1.0 when using RMSE as the criteria: since the α-transformation here converges to RDA, this suggests the compositional constraint impacted the analysis to a lesser extent for this setting. On the other hand, since using MAE produced both lower RMSE and MAE in the validation sets compared with the optimal α determined using RMSE, then we decided to choose the optimal α as 0.9 for the US female data.

4.2 Results: England and Wales data

Using the values of α tuned in Section 4.1, we produced mortality forecasts of proportions of deaths by cause for the test set (England and Wales years 2016–2020 and US years 2012–2021) using the α-transformation coupled with the LC model. We compared this with several LRA

https://doi.org/10.1017/S1748499524000320 Published online by Cambridge University Press

Table 2. Forecast performance on test data, applying CLR, ILR, and the α-transformation coupled with the LC model to England and Wales data from the Human Cause-of-Death Data series (2024), disaggregated for cardiovascular causes of death. For each metric and gender, the bolded values correspond to the error using optimal values of α tuned based on cross-validation. In contrast, underlined values correspond to the lowest metric in the out-of-sample forecast

	RMSE × 100		MAE × 100	
Method	Male	Female	Male	Female
CLR (zeros omitted)	0.1777	0.2125	0.1030	0.1172
CLR (0.25 zero replacement)	0.2311	0.3225	0.1154	0.1740
CLR (0.5 zero replacement)	0.1892	0.2603	0.0980	0.1373
ILR (zeros omitted)	0.1777	0.2125	0.1030	0.1172
ILR (0.25 zero replacement)	0.2311	0.3225	0.1154	0.1740
ILR (0.5 zero replacement)	0.1892	0.2603	0.0980	0.1373
$\alpha = 0.1$	**0.1818**	0.2023	**0.1046**	0.1121
$\alpha = 0.5$	0.1852	0.1714	0.0959	0.1011
$\alpha = 0.7$	0.2109	0.1642	0.1064	0.0994
$\alpha = 0.8$	0.2296	**0.1631**	0.1138	**0.0998**
$\alpha = 0.9$	0.2526	0.1640	0.1228	0.1004
$\alpha = 1$ (RDA)	0.2809	0.1669	0.1329	0.1015

methods in the literature for addressing zeros counts, including the CLR and ILR transformations where zeros were omitted from the data and the CLR and ILR with all zeros replaced by 0.25 or 0.5 before modeling. These additional methods were coupled with an LC model for forecasting and comparison.

Table 2 summarizes the performance of females and males. Aside from the optimal values of α, we also considered values $\alpha = 0.5$, $\alpha = 0.7$, $\alpha = 0.9$, and $\alpha = 1$, the latter equivalent to RDA, that is, ignoring the compositional constraint. The α-transformation, on the whole, tended to produce better forecasting accuracy for the US dataset compared with the CLR and ILR plus either *ad hoc* method of handling zero values. Improvements in the forecast were more evident when assessing MAE across both genders, although even with RMSE, the α-transformation was the second-best performer. Visually, Fig. 3 corroborates the results for males and females, where the α-transformation better fits the observed data when compared with the corresponding CLR and ILR transformations.

The results in Fig. 4 are consistent with the broader observations that overall mortality experienced due to cardiovascular causes in the UK has been improving since the 1960s (British Heart Foundation, 2023; NHS, 2023; Office for National Statistics 2021), although forecasts suggest that an expected decline in the major cardiovascular causes (myocardial infarction and pulmonary heart disease) will be offset by forecast increases in the "other heart" cause category. Again, results from the α-transformation follow the observed data over time more closely compared to CLR and ILR with zeros removed. Moreover, the standard LRA approaches, where a value of 0.25 or 0.5 was added to the zeros, tended to forecast higher proportions for causes with the lowest proportion of deaths (in this case, cardiac arrest), which is offset by lower forecast proportions across all other causes (results shown in Appendix A.3). This result is consistent with the fact that these approaches arbitrarily introduce small death counts where there are none.

4.3 Results: US data

For the larger US cause-of-death dataset, Fig. 5 shows the movement in actual proportion of deaths for each cause over the historical data for US death counts, in a similar way to Fig. 3.

https://doi.org/10.1017/S1748499524000320 Published online by Cambridge University Press

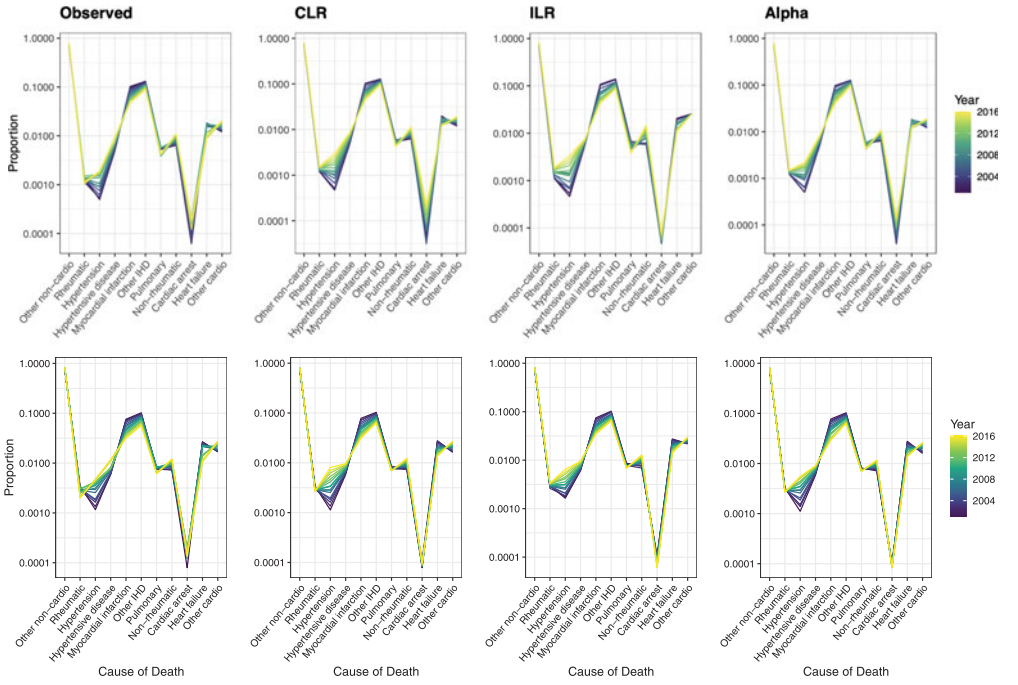


Figure 3. Male (top row) and female (bottom row) mortality by cause in our application to England and Wales data from the Human Cause-of-Death Data series (2024), disaggregated for cardiovascular causes of death. The figures show the movement in actual proportion of deaths for each cause from 2001 to 2016 (left column), while the remaining three columns present results from applying CLR, ILR (with zeros removed), and α-transformations, respectively.

The α-transformation results followed the observed data over time more closely compared to CLR and ILR with zeros removed. This is shown in Table 3 and Fig. 6. Moreover, the standard LRA approaches, where a value of 0.25 or 0.5 was added to the zeros, tended to forecast higher proportions for causes with the lowest proportion of deaths (in this case, cardiac arrest), which is offset by lower forecast proportions across all other causes (results shown in Appendix A.3). This result was consistent with the arbitrary introduction of a small death count where none existed. More importantly, compared with England and Wales data, the forecast performance using the α-transformation was even further improved in the application. Indeed, it suggests that, with a larger volume of data available, our proposed approach can exhibit greater forecasting performance compared to existing log-ratio transformation approaches.

In summary, the point forecast results across both applications suggested that the α-transformation, a generalization of the log-ratio transformation to a broader class of transformations, was an effective way to address zero counts in compositional data, especially compared to *ad hoc* methods of adding small death counts. In Appendix A.4, we performed a sensitivity analysis to assess how much the performance of the two applications depended on the precise α value chosen. Overall, results showed that forecasting performance was largely unaffected when the value of α changed within the tolerance of 0.1 that we employed when tuning this parameter in Section 4.1.

4.4 Interval forecasts

To further understand the projected deaths using the α-transformation, we used interval forecasts to quantify the uncertainty around the point forecast and a further source of (probabilistic) comparison between different methods across both applications of the HCD data (i.e. England and

https://doi.org/10.1017/S1748499524000320 Published online by Cambridge University Press

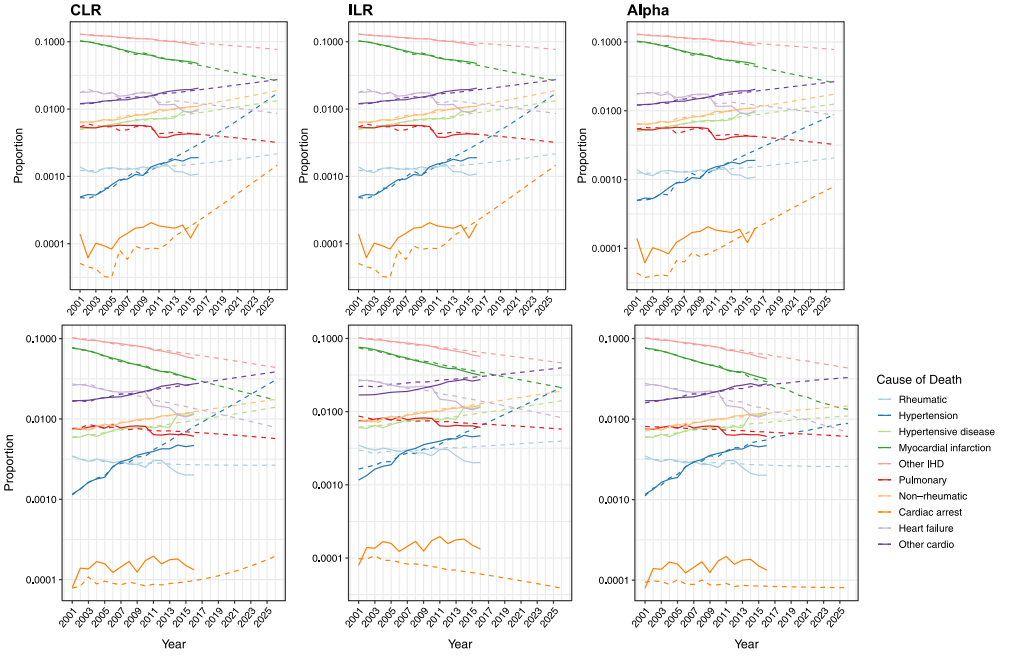


Figure 4. Forecast of cause-specific mortality up to 2026 in our application to England and Wales data from the HCD database, disaggregated for cardiovascular causes of death. Solid lines represent the observed mortality by cause proportions, and dashed lines show the forecast using the CLR, ILR (with zeros removed), and α-transformations (L–R). Mortality by cause is shown for males (top row) and females (bottom row). This figure omits non-cardiovascular causes for presentation purposes.

Wales and US death counts). Briefly, the interval forecasts were produced by adapting the proposed method of Shang and Haberman (2020) for use with the CLR, ILR, and α-transformations and involved the following steps.

(I) Transform the compositional data into the real space using the three methods explored (CLR, ILR, and the proposed α-transformation). Construct the point forecast as per Sections 4.2 and 4.3.

(II) Bootstrap (sample with replacement) the forecast component scores (i.e., $b_{u,c}$ or the age- and cause-specific coefficients that vary over time) and the model fit errors (i.e., $\epsilon_{t,u,c}$) in equation (5). By doing this a large number of times and then taking the empirical quantiles (here, 90% intervals are shown), upper and lower bounds for the interval forecast in real space are produced.

(III) Transform the interval forecast from the real space to the simplex for inference using the corresponding inverse CLR, ILR, or α-transformations. Finally, add back the geometric mean as per the original point estimate approach discussed per equation (5).

Results for the interval forecasts for both applications are presented in Figs. 7 and 8, where the α parameters used in producing interval forecasts were optimized via the interval score approach of Shang and Haberman (2020). Overall, the results across CLR, ILR, and the α-transformation were largely consistent with the corresponding point forecast results shown previously in Figs. 4 and 6. Nevertheless, the interval forecast offers an additional view of uncertainty around the point forecast and reflects the possible extents to which the composition of mortality across different causes could change into the future based on each model.

https://doi.org/10.1017/S1748499524000320 Published online by Cambridge University Press

Table 3. Forecast performance on test data, applying CLR, ILR, and the α-transformation coupled with the LC model to US data from the Human Cause-of-Death Data series (2024), disaggregated for cardiovascular causes of death. For each metric and gender, the bolded values correspond to the error using optimal values of α tuned based on cross-validation. In contrast, underlined values correspond to the lowest metric in the out-of-sample forecast

	RMSE × 100		MAE × 100	
Method	Male	Female	Male	Female
CLR (zeros omitted)	0.3370	0.3819	0.1417	0.1650
CLR (0.25 zero replacement)	0.3566	0.4393	0.1541	0.2049
CLR (0.5 zero replacement)	0.3477	0.4278	0.1467	0.1947
ILR (zeros omitted)	0.3370	0.3819	0.1417	0.1650
ILR (0.25 zero replacement)	0.3566	0.4393	0.1541	0.2049
ILR (0.5 zero replacement)	0.3477	0.4278	0.1467	0.1947
$\alpha = 0.3$	0.3072	0.3138	0.1314	0.1439
$\alpha = 0.5$	0.2905	0.2777	0.1268	0.1300
$\alpha = 0.7$	**0.2877**	0.2518	**0.1299**	0.1202
$\alpha = 0.9$	0.3095	**0.2516**	0.1355	**0.1238**
$\alpha = 1$ (RDA)	0.3435	0.2691	0.1414	0.1287

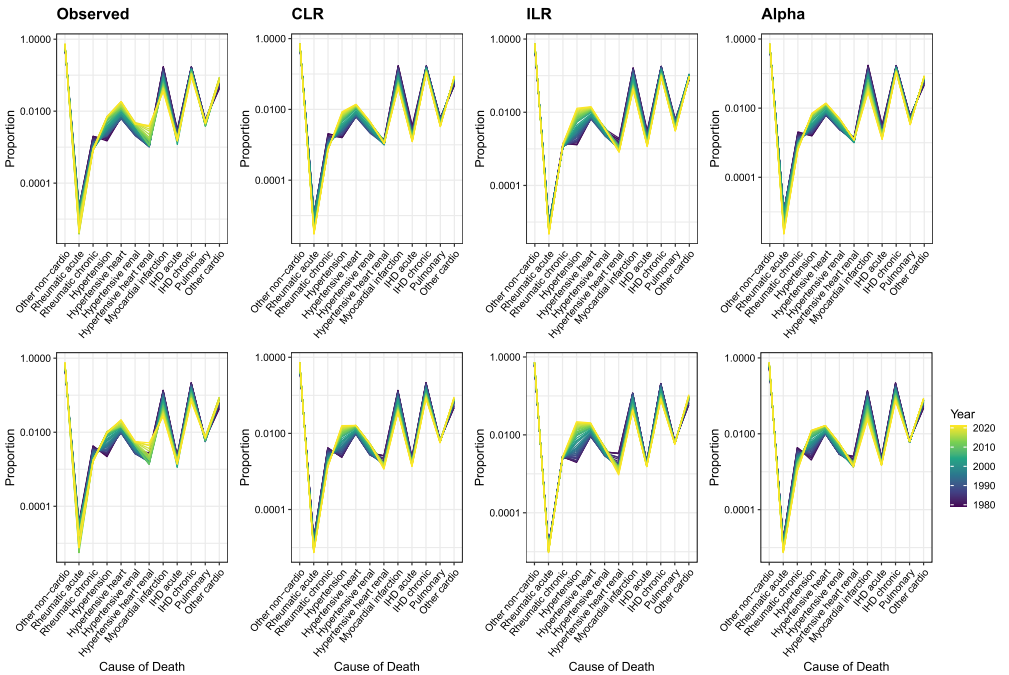


Figure 5. Male (top row) and female (bottom row) mortality by cause in our application to US data from the Human Cause-of-Death Data series (2024), disaggregated for cardiovascular causes of death. The figures show the movement in actual proportion of deaths for each cause from 1979 to 2021 (left column), while the remaining three columns present results from applying CLR, ILR (with zeros removed) and α-transformations, respectively.

4.5 Alternative approaches and future directions

In this section, we consider an alternative approach to the α-transformation for forecasting mortality by cause. Specifically, we consider the multinomial logistic regression (MLR) of Alai *et al.* (2015) and compare its forecast performance to the α-transformation.

https://doi.org/10.1017/S1748499524000320 Published online by Cambridge University Press

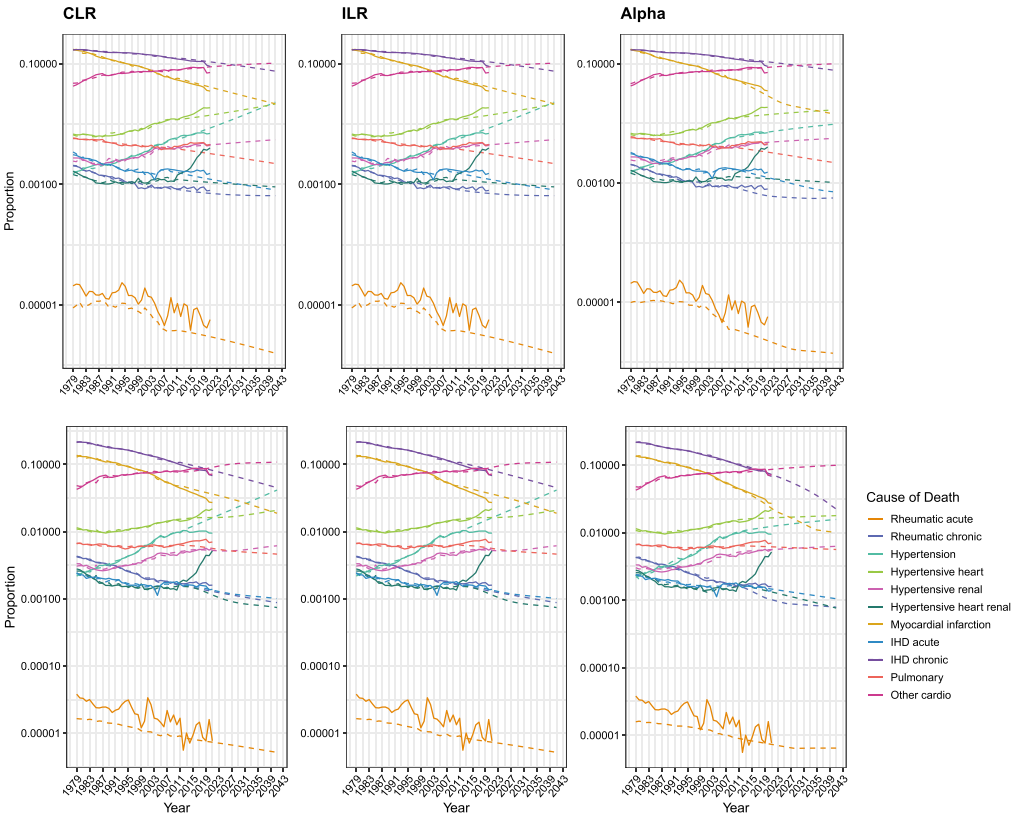


Figure 6. Forecast of cause-specific mortality up to 2051 in our application to US data from the Human Cause-of-Death Data series (2024), disaggregated for cardiovascular causes of death. Solid lines represent the observed mortality by cause proportions, and dashed lines show the forecast using the CLR, ILR (with zeros removed), and α-transformations (L–R). Mortality by cause is shown for males (top row) and females (bottom row). This figure omits non-cardiovascular causes for presentation purposes.

The MLR model is often used to detect factors significantly influencing a response with several competing outcomes. In the literature, numerous applications of the MLR model have been undertaken in cause-of-death analysis over the past three decades. For example, Eberstein *et al.* (1990) used eight categorical and continuous independent variables, including marital status, education, and birth weight, to model five infant cause-specific mortality rates. Lawn *et al.* (2006) applied MLR to model the distribution of neonatal deaths in countries with poor data (see Johnson et al., 2010, for related work). Shahraz *et al.* (2013) employed MLR to redistribute unknown or ill-defined deaths, while Park *et al.* (2006) used it to account for the impact of the tenth revision of the ICD.

For illustrative purposes, we applied the MLR model to US male cause-of-death counts only, disaggregated for cardiovascular causes as per the application in Section 4.3. The forecast performance from applying MLR was assessed using the sum of the squared residual errors. Based on this, we found that the single and simple MLR performed best when compared against the quadratic and cubic MLR. We present results for these in Figs. 9 and 10, which are analogous to those presented earlier in Figs. 5 and 6. Note in assessing the fits, the problem of zeros was still present in the actual death rates by cause; we handled this by adding a 0.01 death count before calculating mortality rates and taking logarithms.

To compare with the forecast performance using CODA methods and shown in Table 3, we calculated the equivalent RMSE and MAE (scaled by 100) for the MLR application to US male

https://doi.org/10.1017/S1748499524000320 Published online by Cambridge University Press

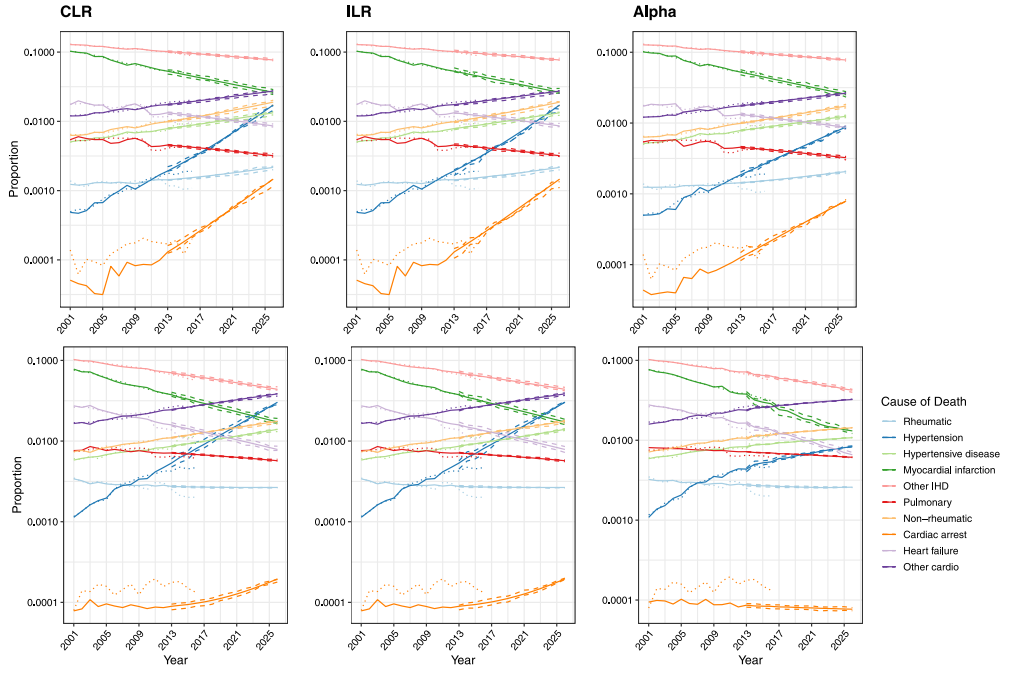


Figure 7. Male (top row) and female (bottom row) 90% interval forecasts up to 2026 in our application to England and Wales data from the Human Cause-of-Death Data series (2024), disaggregated for cardiovascular causes of death.

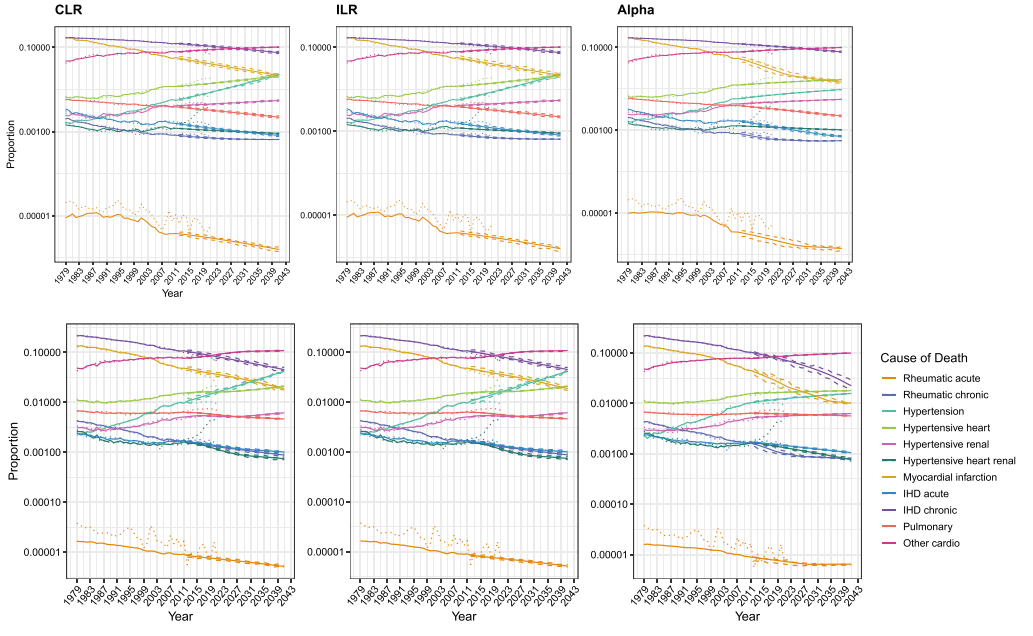


Figure 8. Male (top row) and female (bottom row) 90% interval forecasts up to 2051 in our application to US data from the Human Cause-of-Death Data series (2024), disaggregated for cardiovascular causes of death.

https://doi.org/10.1017/S1748499524000320 Published online by Cambridge University Press

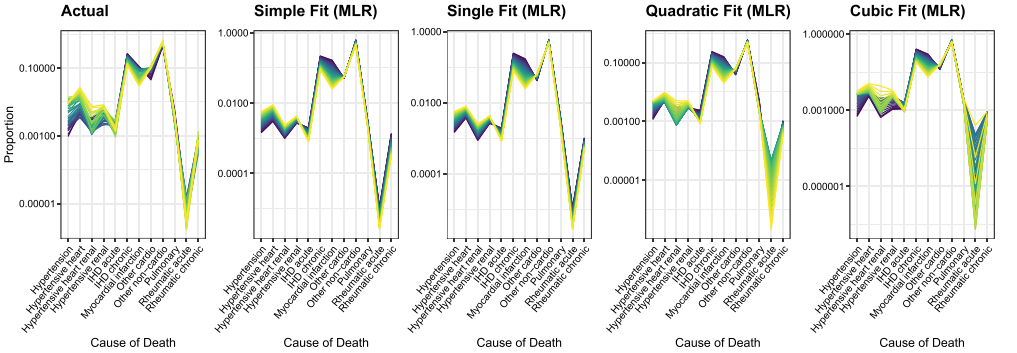


Figure 9. Male mortality by cause using US data from the Human Cause-of-Death Data series (2024), disaggregated for cardiovascular causes of death. The figures show the movement in the actual proportion of deaths for each cause from 1979 to 2021 (left column), while the remaining four columns present results from applying MLR simple, single, quadratic, and cubic regressions, respectively.

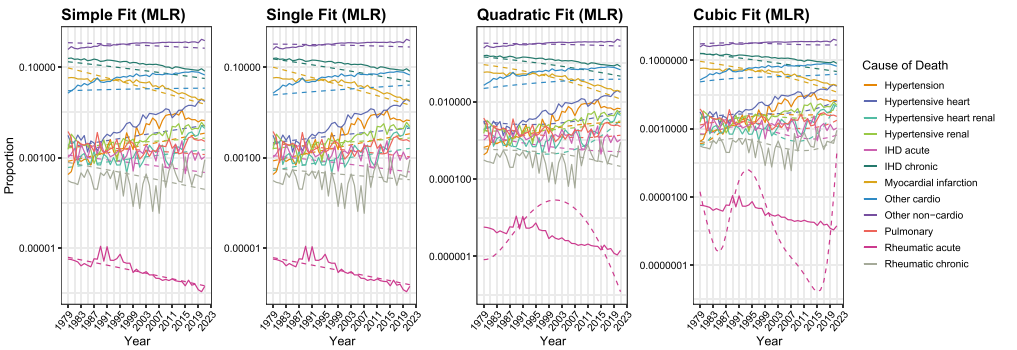


Figure 10. Fits of cause-specific male mortality in our application to US data from the Human Cause-of-Death Data series (2024), disaggregated for cardiovascular causes of death. Solid lines represent the observed mortality by cause proportions, and dashed lines show the fit using MLR regressions. This figure omits non-cardiovascular causes for presentation purposes.

death counts. In this application, the simple MLR produced RMSE and MAE of 2.289 and 1.126, whereas the single MLR produced RMSE and MAE of 2.040 and 1.038. This is substantially higher than the errors of 0.2877 and 0.1299 when we apply the CODA method using an α-transformation. We conjecture similar results would also arise for the case of the US female cause-of-death count data, as well as the England and Wales data. Overall, the comparison indicates that, perhaps not surprisingly, CODA approaches perform better when forecasting using compositional data.

Beyond the MLR model, another method to address the problem of zeros in composition data is applying the Dirichlet distribution. This idea has previously been explored by Tsagris and Stewart (2018) and Graziani and Nigri (2023) in modifying the log-likelihood of the Dirichlet distribution. Such approaches have been applied in other fields, including biology and chromosome detection Tang *et al.* (2022). Further exploration of the Dirichlet composition distribution in understanding mortality by cause would further the understanding of mortality forecasting by cause. Finally, a forecast reconciliation approach can be adopted to ensure forecast coherence instead of applying CODA (Li et al. 2019). Such approaches address the potential problems arising when subpopulation mortality forecasts do not sum up to the aggregate forecast, and they could be considered an alternative approach where there are few zeros in subgroups.

https://doi.org/10.1017/S1748499524000320 Published online by Cambridge University Press

5. Conclusion

In this paper, we have introduced the α-transformation, coupled with an LC model for mortality modeling, as a statistical method to handle cause-of-death compositional data with zero values. Using an expanding window cross-validation approach to select α, we presented two applications to death counts by cause, disaggregated for cardiovascular causes in England and Wales data from 2001 to 2016 and on US data from 1979 to 2021. Forecasts using the α-transformation tend to perform better than those produced using standard log-ratio transformations and is particularly evident in the application to US death counts by cause, having more years of historical data.

We tested a single model (LC) in the compositional framework and focused on heart-related causes of death, where the dataset includes zero counts for several years and age bands. Mortality forecasting by cause may be further enhanced by combining the α-transformation with variations of the LC model, for example, a model which decomposes cause-specific variation into joint and individual variation (Kjaergaard et al., 2019), or using nonparametric techniques such as smoothers or tensor decompositions (Zhang et al., 2023). Also, rather than adding a small death count or removing zeros entirely, other approaches could be compared against the α-transformation, including "borrowing" from a neighboring age (for the same cause) or smoothing over similar causes (for the same age), along with other imputation methods (Lubbe et al., 2021). We leave such investigations as avenues for future study.

One feature of the death counts by cause for both England and Wales and the USA, which is true of many other cause-of-death datasets in other countries, is that zero counts of death for multiple causes tend to occur across consecutive years and/or adjacent age groups. In other settings with fewer or no zeros count, and where the occurrence of the zeros is more sporadic, simpler approaches, such as adding a small value to enable LRA may have fewer implications on the analysis and conclusions relative to using the α-transformation (Tsagris & Stewart, 2022). Conversely, for older ages and emerging causes with only recent data (including COVID-19), reflecting true zeros in the data in a statistically more rigorous and data-driven manner, as the α-transformation does, is expected to produce more accurate forecasts.

While CODA is useful in capturing dependencies between causes arising due to the compositional nature of the data, other dependencies, such as comorbidities, can arise irrespective of how the data are treated. An important avenue of future research is how methods such as α-transformation could be coupled with techniques that can account for such dependencies. Indeed, an essential application of CODA for life insurers is to enhance the understanding of morbidity and mortality risks. CODA can also be used to investigate the risk implications across different subgroups of insured lives and exposures, and we anticipate the α-transformation will play a useful role in modeling compositional data arising from these other settings. Finally, the results from the application suggest that while the aggregate cardiovascular death counts are expected to reduce, some granular causes of death within the cardiovascular cause are expected to increase, particularly for males across England and Wales. Analysis using US death counts indicate slight decreases across all granular cardiovascular causes. These findings should be further investigated, along with other causes.

Acknowledgments. The authors are grateful for the comments from the participants at the Insurance Data Science conference in 2023 and the Conference in Celebration of David Wilkie's 90th birthday in 2024.

Funding statement. The second author was supported by an Australian Research Council Discovery Project, DP230102250, and the third author was supported by an Australian Research Council Discovery Project, DP230101908.

Data availability statement. The data and code that support the findings of this study are openly available at https://github.com/zm-dong/coda_cause_mortality.

https://doi.org/10.1017/S1748499524000320 Published online by Cambridge University Press

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, **44**(2), 139–160.

Alai, D. H., Arnold, S., & **Sherris, M.** (2015). Modelling cause-of-death mortality and the impact of cause-elimination. *Annals of Actuarial Science*, **9**(1), 167–186.

Arnold, S., & **Sherris, M.** (2013). Forecasting mortality trends allowing for cause-of-death mortality dependence. *North American Actuarial Journal*, **17**(4), 273–282.

Basel Committee on Banking Supervision. (2013). Longevity risk transfer markets: Market structure, growth drivers and impediments, and potential risks. https://www.bis.org/publ/joint34.pdf.

Bergeron-Boucher, M.-P., Canudas-Romo, V., Oeppen, J., & **Vaupel, J. W.** (2017). Coherent forecasts of mortality with compositional data analysis. *Demographic Research*, **37**, 527–566.

Bergeron-Boucher, M.-P., & **Kjærgaard, S.** (2022). Mortality forecasting at age 65 and above: An age-specific evaluation of the Lee-Carter model. *Scandinavian Actuarial Journal*, **2022**(1), 64–79.

Bergeron-Boucher, M.-P., Strozza, C., Simonacci, V., & **Oeppen, J.** (2022). Modeling and forecasting healthy life expectancy with compositional data analysis. SocArXiv. July, 9.

Booth, H., Maindonald, J., & **Smith, L.** (2002). Applying Lee-Carter under conditions of variable mortality decline. *Population Studies*, **56**(3), 325–336.

Booth, H., Tickle, L., & **Smith, L.** (2005). Evaluation of the variants of the Lee-Carter method of forecasting mortality: A multi-country comparison. *New Zealand Population Review*, **31**(1), 13–34.

Box, G. E., & **Cox, D. R.** (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, **26**(2), 211–243.

British Heart Foundation. (2023). UK Factsheet. https://www.bhf.org.uk/-/media/files/for-professionals/research/heart-statistics/bhf-cvd-statistics-uk-factsheet.pdf.

Cairns, A. J., Blake, D., & **Dowd, K.** (2006). A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. *Journal of Risk and Insurance*, **73**(4), 687–718.

Chai, T., & **Draxler, R. R.** (2014). Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, **7**(3), 1247–1250.

Eberstein, I. W., Nam, C. B., & **Hummer, R. A.** (1990). Infant mortality by cause of death: Main and interaction effects. *Demography*, **27**(3), 413–430.

Gao, G., & **Shi, Y.** (2021). Age-coherent extensions of the Lee–Carter model. *Scandinavian Actuarial Journal*, **2021**(10), 998–1016.

Graziani, R., & **Nigri, A.** (2023). An age-period-cohort model in a dirichlet framework: A coherent causes of death estimation. Technical report, SocArXiv. https://ideas.repec.org/p/osf/socarx/856yw.html.

Greenacre, M. (2021). Compositional data analysis. *Annual Review of Statistics and its Application*, **8**(1), 271–299.

Greenacre, M. (2024). The chipower transformation: A valid alternative to logratio transformations in compositional data analysis. *Advances in Data Analysis and Classification*, **18**(3), 769–796, Working paper, arXiv.

Greenacre, M., & **Grunsky, E.** (2019). The isometric logratio transformation in compositional data analysis: A practical evaluation. Working paper 1627, Universitat Pompeu Fabra.

Grifoll, M., Ortego, M., & **Egozcue, J.** (2019). Compositional data techniques for the analysis of the container traffic share in a multi-port region. *European Transport Research Review*, **11**(1), 1–15.

Hodson, T. O. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geoscientific Model Development*, **15**(14), 5481–5487.

Holford, T. R. (1983). The estimation of age, period and cohort effects for vital rates. *Biometrics*, **39**(2), 311–324.

Human Cause-of-death Data series (2024). French Institute for Demographic Studies (France), Max Planck Institute for Demographic Research (Germany) and the University of California, Berkeley (USA),

Hyndman, R. J., Booth, H., & **Yasmeen, F.** (2013). Coherent mortality forecasting: the product-ratio method with functional time series models. *Demography*, **50**(1), 261–283.

Johnson, H. L., Liu, L., Fischer-Walker, C., & **Black, R. E.** (2010). Estimating the distribution of causes of death among children age 1-59 months in high-mortality countries with incomplete death certification. *International Journal of Epidemiology*, **39**(4), 1103–1114.

Kjaergaard, S., Ergemen, Y. E., Bergeron-Boucher, M.-P., Oeppen, J., & **Kallestrup-Lamb, M.** (2020). Longevity forecasting by socio-economic groups using compositional data analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **183**(3), 1167–1187.

Kjaergaard, S., Ergemen, Y. E., Kallestrup-Lamb, M., Oeppen, J., & **Lindahl-Jacobsen, R.** (2019). Forecasting causes of death using compositional data analysis: The case of cancer deaths. *Journal of the Royal Statistical Society Series C: Applied Statistics*, **68**(5), 1351–1370.

Lawn, J. E., Wilczynska-Ketende, K., & **Cousens, S. N.** (2006). Estimating the causes of 4 million neonatal deaths in the year 2000. *International Journal of Epidemiology*, **35**(3), 706–718.

Lee, R., & **Miller, T.** (2001). Evaluating the performance of the lee-carter method for forecasting mortality. *Demography*, **38**(4), 537–549.

https://doi.org/10.1017/S1748499524000320 Published online by Cambridge University Press

Lee, **R. D.**, & **Carter**, **L. R.** (1992). Modeling and forecasting US mortality. *Journal of the American Statistical Association: Applications & Case Studies*, **87**(419), 659–671.
Li, **H.**, **Li**, **H.**, **Lu**, **Y.**, & **Panagiotelis**, **A.** (2019). A forecast reconciliation approach to cause-of-death mortality modeling. *Insurance: Mathematics and Economics*, **86**, 122–133.
Li, **H.**, & **Lu**, **Y.** (2019). Modeling cause-of-death mortality using hierarchical archimedean copula. *Scandinavian Actuarial Journal*, **2019**(3), 247–272.
Li, **N.**, & **Lee**, **R.** (2005). Coherent mortality forecasts for a group of populations: An extension of the lee-carter method. *Demography*, **42**(3), 575–594.
Lubbe, **S.**, **Filzmoser**, **P.**, & **Templ**, **M.** (2021). Comparison of zero replacement strategies for compositional data with large numbers of zeros. *Chemometrics and Intelligent Laboratory Systems*, **210**, 104248.
Martin-Fernandez, **J. A.**, **Barcelo-Vidal**, **C.**, & **Pawlowsky-Glahn**, **V.** (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, **35**(3), 253–278.
National Institute for Health and Care Excellence (2023). CVD risk assessment and management: What is the impact of CVD? https://cks.nice.org.uk/topics/cvd-risk-assessment-management/background-information/burden-of-cvd/.
NHS. (2023). Cardiovascular disease. https://www.nhs.uk/conditions/cardiovascular-disease/.
Oeppen, **J.** (2008). Coherent forecasting of multiple-decrement life tables: A test using Japanese cause of death data.
Office for National Statistics (2021). Ischaemic heart diseases deaths including comorbidities, England and Wales: 2019 registrations, https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/ischaemicheartdiseasesdeathsincludingcomorbiditiesenglandandwales/2019registrations.
Park, **Y.**, **Choi**, **J. W.**, & **Lee**, **D.-H.** (2006). A parametric approach for measuring the effect of the 10th revision of the international classification of diseases. *Journal of the Royal Statistical Society: Series C*, **55**(5), 677–697.
Racine, **J.** (2000). Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of Econometrics*, **99**(1), 39–61.
Raleigh, **V.**, **Jefferies**, **D.**, & **Wellings**, **D.** (2022). Cardiovascular diseases in England: Supporting leaders to take action. https://www.kingsfund.org.uk/publications/cardiovascular-disease-england.
Renshaw, **A. E.**, & **Haberman**, **S.** (2003). Lee–Carter mortality forecasting with age-specific enhancement. *Insurance: Mathematics and Economics*, **33**(2), 255–272.
Renshaw, **A. E.**, & **Haberman**, **S.** (2006). A cohort-based extension to the Lee-Carter model for mortality reduction factors. *Insurance: Mathematics and economics*, **38**(3), 556–570.
Schnaubelt, **M.** (2019). A comparison of machine learning model validation schemes for non-stationary time series data. Technical report, FAU Discussion Papers in Economics.
Shahraz, **S.**, **Bhalla**, **K.**, **Lozano**, **R.**, **Bartels**, **D.**, & **Murray**, **C. J. L.** (2013). Improving the quality of road injury statistics by using regression models to redistribute ill-defined events. *Injury Prevention*, **19**(1), 1–5.
Shang, **H. L.**, & **Haberman**, **S.** (2020). Forecasting age distribution of death counts: An application to annuity pricing. *Annals of Actuarial Science*, **14**(1), 150–169.
Tang, **M.-L.**, **Wu**, **Q.**, **Yang**, **S.**, & **Tian**, **G.-L.** (2022). Dirichlet composition distribution for compositional data with zero components: An application to fluorescence in situ hybridization (fish) detection of chromosome. *Biometrical Journal*, **64**(4), 714–732.
Tsagris, **M.**, & **Stewart**, **C.** (2018). A dirichlet regression model for compositional data with zeros. *Lobachevskii Journal of Mathematics*, **39**(3), 398–412.
Tsagris, **M.**, & **Stewart**, **C.** (2020). A folded model for compositional data analysis. *Australian & New Zealand Journal of Statistics*, **62**(2), 249–277.
Tsagris, **M.**, & **Stewart**, **C.** (2022). A review of flexible transformations for modeling compositional data. In W. He, L. Wang, J. Chen & C. D. Lin (Eds.), *Advances and innovations in statistics and data science* (pp. 225–234). Springer.
Tsagris, **M. T.**, **Preston**, **S.**, & **Wood**, **A. T.** (2011). A data-based power transformation for compositional data. In *Proceedings of the 4th international workshop on compositional data analysis, Girona, Spain*, https://arxiv.org/abs/1106.1451.
World Health Organization. (1992). The ICD-10 classification of mental and behavioural disorders. https://cdn.who.int/media/docs/default-source/classification/other-classifications/9241544228_eng.pdf.
Zhang, **X.**, **Huang**, **F.**, **Hui**, **F. K. C.**, & **Haberman**, **S.** (2023). Cause-of-death mortality forecasting using adaptive penalized tensor decompositions. *Insurance: Mathematics and Economics*, **111**, 193–213.

A. Supplementary Information and Results

A.1 The Lee-Carter model for modeling mortality

This paper applies the LC mortality model after LRA and the α-transformation (Lee & Carter, 1992). For completeness, this appendix provides a brief review of the LC model for analysis of non-compositional data, that is, RDA.

https://doi.org/10.1017/S1748499524000320 Published online by Cambridge University Press

Treating causes independently, the LC model fits and predicts central mortality rates by expressing the log mortality rate as a linear function of a time factor with age parameters. For cause c, let $m_{t,u,c}$ denote the central death rate for age u in year t, which we compute as $m_{t,u,c} = d_{t,u,c}/L_{t,u}$, where the denominator $L_{t,u}$ is the exposure of person-years lived at age u. The LC model is then defined as

$$\ln(m_{t,u,c}) = \mu_{u,c} + b_{u,c}k_{t,c} + \epsilon_{t,u,c}, \tag{A1}$$

where $\mu_{u,c}$ represents an age- and cause-specific average mortality over time, $b_{u,c}$ denotes the age- and cause-specific coefficients that vary over time; $k_{t,c}$ denotes a factor of time-varying indices for the level of mortality, and the $\epsilon_{t,u,c}$ denote residual error terms. The model is typically fitted by applying a singular value decomposition to a $U \times T$ matrix the elements of which are given by $\ln(m_{t,u,c})$, after subtracting the average mortality rate over time for a given cause. After fitting, mortality forecasting is performed by modeling the estimated time factors $k_{t,i}$ as an autoregressive integrated moving average time series. The common choice is a simple random walk with drift. We refer the reader to Lee and Carter (1992) for more details regarding parameter estimation of the LC model.

The LC model is commonly used for national forecasts, with its primary advantages including its simplicity, ability to deal with uncertainty, and low requirement for subjective judgment (Bergeron-Boucher & Kjærgaard, 2022). With its simplicity comes a number of limitations, and consequently many variations of LC exist to improve its performance. Among many others, examples include Renshaw and Haberman (2003), which generalized the LC model to include more than one factor; the Cairns *et al.* (2006) model, which is a popular alternative that models the probability of survival rather than the $\log_{10}$ mortality rates; the Lee and Miller (2001) and Booth *et al.* (2002) models, both of which aim to improve the forecasting performance of the LC model (Booth et al., 2005); and Li and Lee (2005) and Gao and Shi (2021), who apply coherence in the context of mortality modeling and age-coherent extensions of LC, respectively.

A.2 Additional results for the application to the HCD database

Table A.1 shows the results from cross-validation for England and Wales's cause-of-death data. Based on cross-validation, we determined the optimal α value is 0.1 for males and 0.8 for females. This was then applied to produce the results in Section 4.2.

Table A.2 similarly shows the results from cross-validation for US death counts by cause to determine the optimal α. Based on cross-validation, we determined the optimal α value is 0.7 for males and 0.9 for females. This was applied to produce the results in Section 4.3.

A.3 Additional results comparing forecast performance using CLR and ILR transformations with different techniques to replace zero counts

We further compared the performance of CLR and ILR forecasts when zero counts are replaced by 0.25 or 0.5 for both England and Wales and US death counts by cause of death. This was applied for both male and female death counts on both sets of data for completeness. These results are included in Sections 4.2 and 4.3. For England and Wales, Figs. A.1 and A.2 show the visualizations of the forecasts when different zero replacement approaches are used. The forecast and trends change and are sensitive to the method of zero replacement. The α-transformation presents a statistical approach that removes this sensitivity.

Similarly, for US death counts, Figs. A.3 and A.4 show visualizations of the forecasts when different zero replacement approaches are used. It is worth noting that longer term trends are still impacted by different approaches to replace zeros, despite the US dataset having a longer history compared to the England and Wales death counts by cause.

https://doi.org/10.1017/S1748499524000320 Published online by Cambridge University Press

Table A.1 Results for validation sets (RMSE and MAE, based on fourfold expanding window cross-validation) to tune α, using the α-transformation coupled with an LC model for forecasting in our application to England and Wales death counts by cause from the Human Cause-of-Death Data series (2024), disaggregated for cardiovascular causes of death. Optimal values of α are shown in bold, noting all results are scaled by multiplying by 100

	RMSE		MAE	
α	Male	Female	Male	Female
0 (CLR)	0.1919	0.2022	0.0985	0.0931
0 (ILR)	0.1919	0.2022	0.0985	0.0931
0.1	**0.1992**	0.2001	**0.0766**	0.0727
0.2	0.2037	0.1903	0.0791	0.0694
0.3	0.2099	0.1813	0.0812	0.0669
0.4	0.2542	0.1733	0.0924	0.0649
0.5	0.2641	0.1660	0.0961	0.0633
0.6	0.2757	0.1595	0.1000	0.0619
0.7	0.2882	0.1539	0.1043	0.0607
0.8	0.3200	0.1500	0.1135	**0.0602**
0.9	0.3347	**0.1492**	0.1182	0.0613
1 (RDA)	0.3327	0.1517	0.1174	0.0632

Table A.2 Results for validation sets (RMSE and MAE, based on tenfold expanding window cross-validation) to tune α, using the α-transformation coupled with an LC model for forecasting in our application to US death counts by cause from the Human Cause-of-Death Data series (2024), disaggregated for cardiovascular causes of death. Optimal values of α are shown in bold, noting all results are scaled by multiplying by 100

	RMSE		MAE	
α	Male	Female	Male	Female
0 (CLR)	0.2320	0.3078	0.1092	0.1195
0 (ILR)	0.2320	0.3078	0.1092	0.1195
0.1	0.2244	0.3101	0.0827	0.0963
0.2	0.2146	0.2964	0.0797	0.0929
0.3	0.2061	0.2843	0.0771	0.0898
0.4	0.1990	0.2736	0.0750	0.0871
0.5	0.1933	0.2648	0.0732	0.0848
0.6	0.1891	0.2560	0.0717	0.0827
0.7	**0.1868**	0.2485	**0.0709**	0.0810
0.8	0.1871	0.2418	0.0709	0.0794
0.9	0.1913	0.2390	0.0721	**0.0788**
1 (RDA)	0.1998	**0.2386**	0.0744	0.0791

A.4 Sensitivity analysis of the choice of α

A sensible question to ask then is how sensitive were the results to the particular chosen values of α as long as we were within this tolerance range. Based on additional testing, we found that results remain largely unaffected when α was specified within 0.1.

The optimal α for England and Wales death counts (Section 4.2) is 0.1 for males, resulting in RMSE and MAE of 0.1818 and 0.1046, respectively. For $\alpha = 0.09$, the resulting RMSE and MAE are 0.1832 and 0.1055. For $\alpha = 0.11$, the resulting RMSE and MAE are 0.1806 and 0.1037. Here, the results improve when $\alpha = 0.11$, compared to specifying α to the nearest 0.1. However, the resulting inferences around mortality forecasts by cause are unchanged.

https://doi.org/10.1017/S1748499524000320 Published online by Cambridge University Press

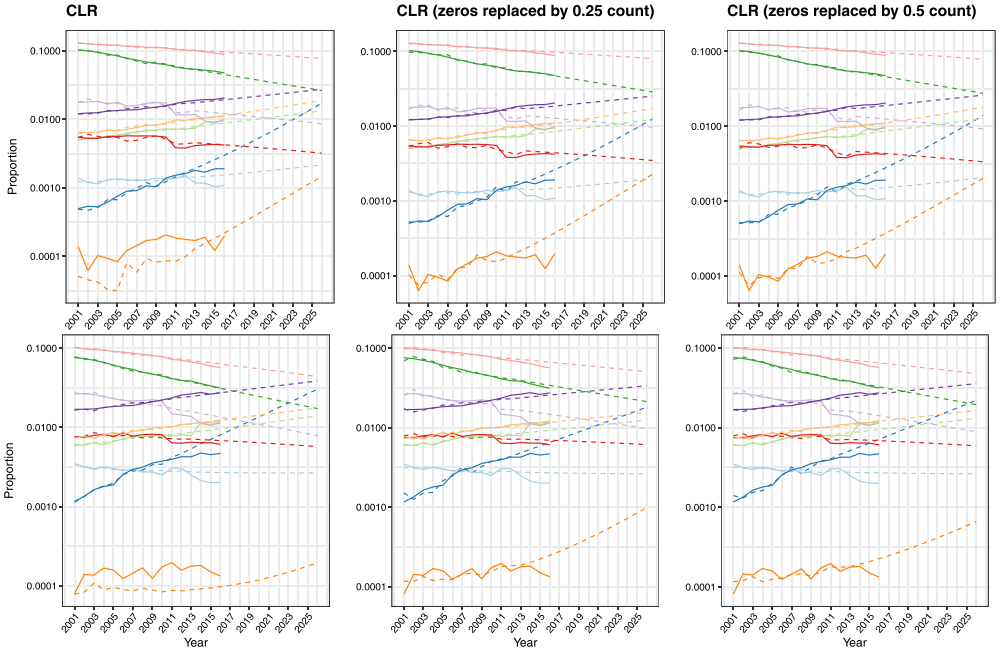


Figure A.1. Forecast of cause-specific mortality up to 2026 in our application to England and Wales death counts by cause from the Human Cause-of-Death Data series (2024), disaggregated for cardiovascular causes of death. Solid lines represent the observed mortality by cause proportions, and dashed lines show the forecast using the CLR transformation with variations in the treatment of zeros in the data. Mortality by cause is shown for males (top row) and females (bottom row). This figure omits non-cardiovascular causes for presentation purposes.

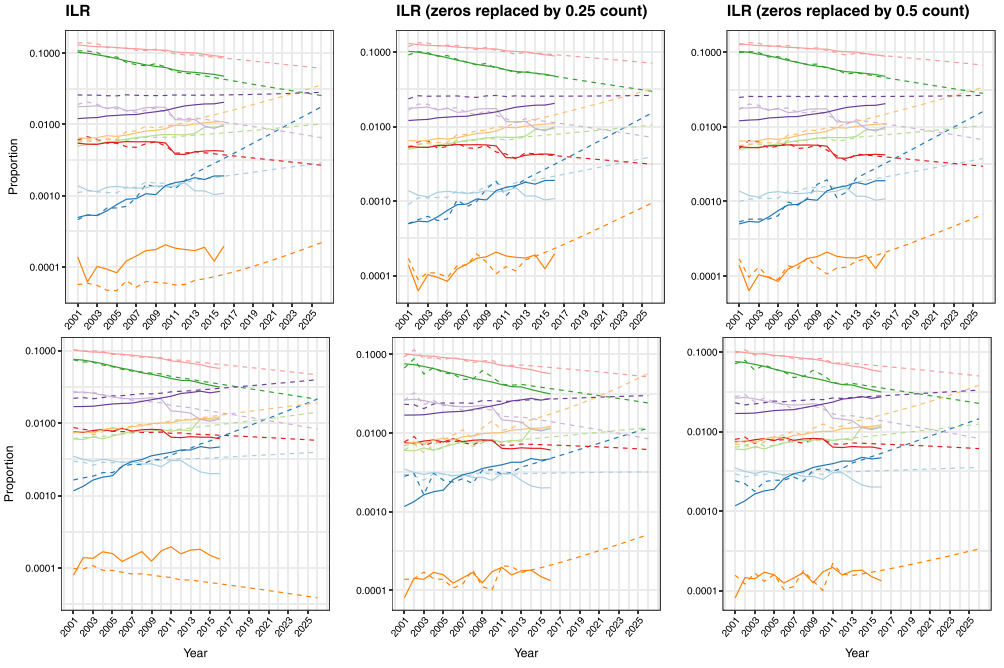


Figure A.2. Forecast of cause-specific mortality up to 2026 in our application to England and Wales death counts by cause from the Human Cause-of-Death Data series (2024), disaggregated for cardiovascular causes of death. Solid lines represent the observed mortality by cause proportions, and dashed lines show the forecast using the ILR transformation with variations in the treatment of zeros in the data. Mortality by cause is shown for males (top row) and females (bottom row). This figure omits non-cardiovascular causes for presentation purposes.

https://doi.org/10.1017/S1748499524000320 Published online by Cambridge University Press

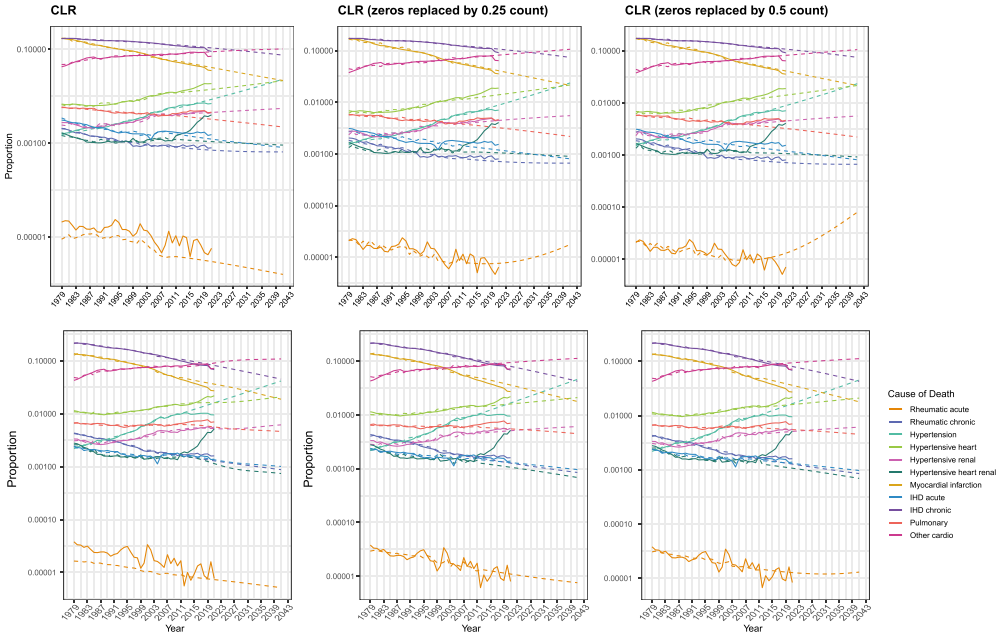


Figure A.3. Forecast of cause-specific mortality up to 2051 in our application to US death counts by cause from the Human Cause-of-Death Data series (2024), disaggregated for cardiovascular causes of death. Solid lines represent the observed mortality by cause proportions, and dashed lines show the forecast using the CLR transformation with variations in the treatment of zeros in the data. Mortality by cause is shown for males (top row) and females (bottom row). This figure omits non-cardiovascular causes for presentation purposes.

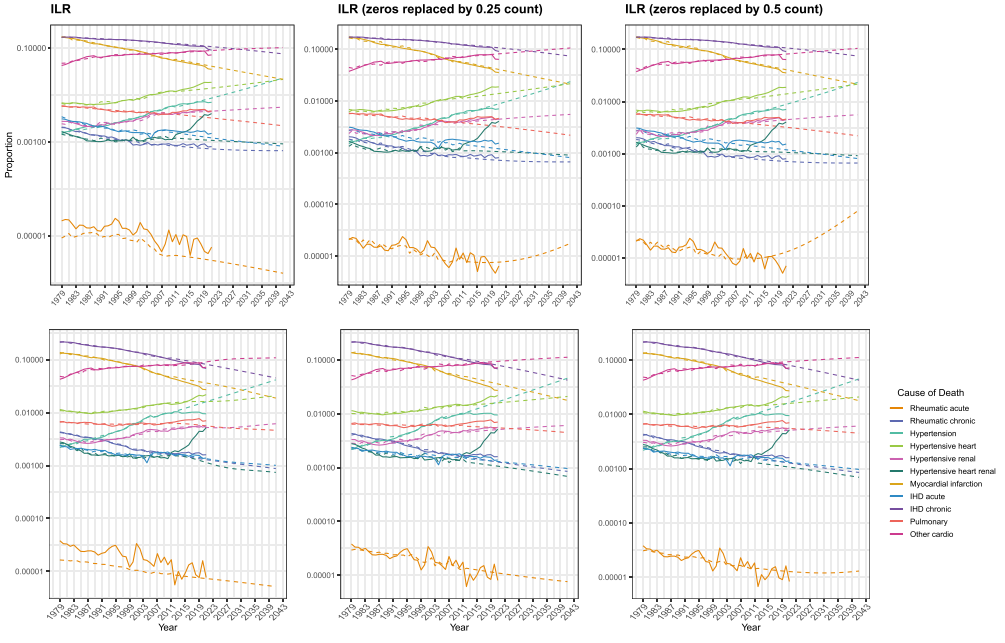


Figure A.4. Forecast of cause-specific mortality up to 2051 in our application to US death counts by cause from the Human Cause-of-Death Data series (2024), disaggregated for cardiovascular causes of death. Solid lines represent the observed mortality by cause proportions, and dashed lines show the forecast using the ILR transformation with variations in the treatment of zeros in the data. Mortality by cause is shown for males (top row) and females (bottom row). This figure omits non-cardiovascular causes for presentation purposes.

https://doi.org/10.1017/S1748499524000320 Published online by Cambridge University Press

We perform a similar exercise on the optimal alphas for US data, where there is a longer history of death counts. For example, the optimal α for US females (Section 4.3) is 0.9, resulting in RMSE and MAE of 0.2516 and 0.1238, respectively. For $\alpha = 0.91$, the resulting RMSE and MAE are 0.2528 and 0.1243. For $\alpha = 0.89$, the resulting RMSE and MAE are 0.2504 and 0.1233, an improvement to the selected optimal $\alpha = 0.90$. Moreover, the resulting inferences from the forecast were largely unchanged in terms of shape and trend in the forecast of cause-specific mortality.

Cite this article: Dong ZM, Shang HL, Hui F and Bruhn A (2025). A compositional approach to modeling cause-specific mortality with zero counts, *Annals of Actuarial Science*, **19**, 416–441. https://doi.org/10.1017/S1748499524000320

https://doi.org/10.1017/S1748499524000320 Published online by Cambridge University Press