The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1035–1050
https://doi.org/10.1214/23-AOAS1821
© Institute of Mathematical Statistics, 2024

INFORMATION-INCORPORATED CLUSTERING ANALYSIS OF DISEASE PREVALENCE TRENDS

BY CHENJIN MA[1,a], CUNJIE LIN[2,b], YUAN XUE[3,c], SANGUO ZHANG[3,d], QINGZHAO ZHANG[4,e] AND SHUANGGE MA[5,f]

[1]*Department of Statistics and Data Science, Beijing University of Technology,* [a]*machenjin@bjut.edu.cn*
[2]*School of Statistics, Renmin University of China,* [b]*lincunjie@ruc.edu.cn*
[3]*School of Mathematics Sciences, University of Chinese Academy of Sciences,* [c]*xueyuan115@mails.ucas.ac.cn,* [d]*sgzhang@ucas.ac.cn*
[4]*Department of Statistics, School of Economics, Xiamen University,* [e]*qzzhang@xmu.edu.cn*
[5]*Department of Biostatistics, Yale School of Public Health,* [f]*shuangge.ma@yale.edu*

In biomedical research the analysis of disease prevalence is of critical importance. While most of the existing prevalence studies focus on individual diseases, there has been increasing effort that jointly examines the prevalence values and their trends of multiple diseases. Such joint analysis can provide valuable insights not shared by individual-disease analysis. A critical limitation of the existing analysis is that there is a lack of attention to existing information, which has been accumulated through a large number of studies and can be valuable especially when there are a large number of diseases but the number of prevalence values for a specific disease is limited. In this study we conduct the functional clustering analysis of prevalence trends for a large number of diseases. A novel approach based on the penalized fusion technique is developed to incorporate information mined from published articles. It is innovatively designed to take into account that such information may not be fully relevant or correct. Another significant development is that statistical properties are rigorously established. Simulation is conducted and demonstrates its competitive performance. In the analysis of data from Taiwan NHIRD (National Health Insurance Research Database), new and interesting findings that differ from the existing ones are made.

1. Introduction. In biomedical research the analysis of prevalence has had a pivotal role. There have been extensive studies on the "snapshot" values of disease prevalence and their associated factors. Quite a few other studies are on disease prevalence trends, and such analysis can assist prioritizing diseases (e.g., those with fast increasing prevalence should receive more attention), identifying new risk factors for etiology (which can facilitate developing prevention and treatment strategies), and managing diseases in clinical practice (e.g., by making proper resource planning and allocation). Most of the existing prevalence studies focus on individual diseases (or small classes of preselected, tightly connected diseases). On the other hand, there has been increasing effort that jointly analyzes the prevalence values of two or more diseases (Joffres et al. (2013), Romanowski et al. (2015), Jadhav et al. (2021)). For example, a prospective observational study conducted in the U.S. examines the prevalence of chest pain and acute myocardial infarction (MI) and shows that patients without chest pain on presentation represent a large segment of the MI population and have an increased risk for delays in seeking treatment (Canto et al. (2000)). Another example is the examination of the prevalence values and trends of multiple cancers under the metastasis networks (Chen et al. (2009)). Also, using the Taiwan NHI (National Health Insurance) data, studies have been conducted on diseases sharing similar prevalence trends with HIV/AIDS (Lai (2015)) and amyotrophic lateral sclerosis (Tsai, Hu and Lee (2019)).

Received September 2022; revised June 2023.
Key words and phrases. Clustering, disease prevalence trends, information incorporation, penalized fusion.

The growth in joint prevalence analysis fits the paradigm shifting in biomedical research from individual-disease to pan-disease analysis. One of the early breakthroughs is the human disease network (HDN) analysis, which examines the interconnections among diseases based on their genetic risk factors (Goh et al. (2007)). The phenotypic HDN (pHDN) analysis has been subsequently conducted and differs from the molecular HDN analysis by focusing on clinical phenotypes (Zhou et al. (2014)). With the consideration that both molecular basis and disease phenotypes are not "close enough" to clinics, pan-disease analysis of the interconnections between disease clinical treatment measures, such as treatment cost (Ma et al. (2020)) and inpatient length of stay, has been conducted. Here we note that disease prevalence trend is correlated with the aforementioned variables (e.g., a disease with low prevalence is likely to have limited treatment cost and inpatient stay) but cannot be fully derived from them. As such, this work is expected to complement but not strongly overlap with the aforementioned ones.

In this study we examine the interrelationships among diseases in terms of prevalence trend. To fix ideas, in the upper-left panel of Figure 1, we present the prevalence trends of 10 diseases. In the upper-right panel, we move the curves vertically and find four clusters, with those in the same cluster having similar patterns. Although functional clustering has been extensively conducted (Jacques and Preda (2014)), its application to disease prevalence trends has been very limited but can have important implications. First, the temporal variations of disease prevalence can be largely attributed to the development of prevention programs, improvement in diagnosis, change in environmental conditions, and other time-dependent influential factors. As such, if multiple diseases have similar temporal trends, it is sensible to hypothesize that they share time-dependent risk factors and/or are affected by similar prevention/diagnosis/treatment programs. Identifying such shared factors can advance our understanding of diseases and inform clinical practice. For example, epidemiologic studies suggest that cognitive dysfunction and type 2 diabetes are "correlated," which is manifested in the shared patterns of their prevalence trends (Luchsinger et al. (2007)). Motivated by such observations, researchers have examined the micro-relation of glycemic status with different domains of cognitive functions and found a number of vascular and neurodegenetive mechanisms through which type 2 diabetes and cognitive function are interconnected. Second, clustering analysis can lead to a new/alternative way of disease classification and characterization. Different classifications are needed to serve different purposes (Zhou et al. (2018)), and the classification, as exemplified in the upper-right panel of Figure 1, has a basis different from those based on organ, symptom, and genetics. It can be "closer" to public health as well as medical care management and planning. Third, clustering analysis can lead to simpler data structures and more accurate estimation, as estimation only needs to be done at the cluster (as opposed to individual) level.

Among the existing studies, the most relevant is Jadhav et al. (2021), which also conducts the functional clustering analysis of disease prevalence trends and analyzes the NHI data. The present study advances from Jadhav et al. (2021) in multiple important aspects. As exemplified in the upper-left panel of Figure 1, for each disease the number of measurements (usually one prevalence value per year) is limited. Combined with the large number of diseases, this leads to a lack of information and hence unreliable estimation and unsatisfactory clustering. Many diseases have been extensively studied in published literature, and quite a few studies have touched on the "interconnections" among diseases. In Figure 2, for selected diseases, we present the numbers of published studies that have mentioned disease pairs (more details on information extraction is presented in Section 2.2). For example, our information extraction identifies 1431 publications that have mentioned both bipolar and dementias. Information as sketched in Figure 2 has been accumulated through a large number of studies and can be valuable. *Methodologically, this study significantly advances from Jadhav et al. (2021)*

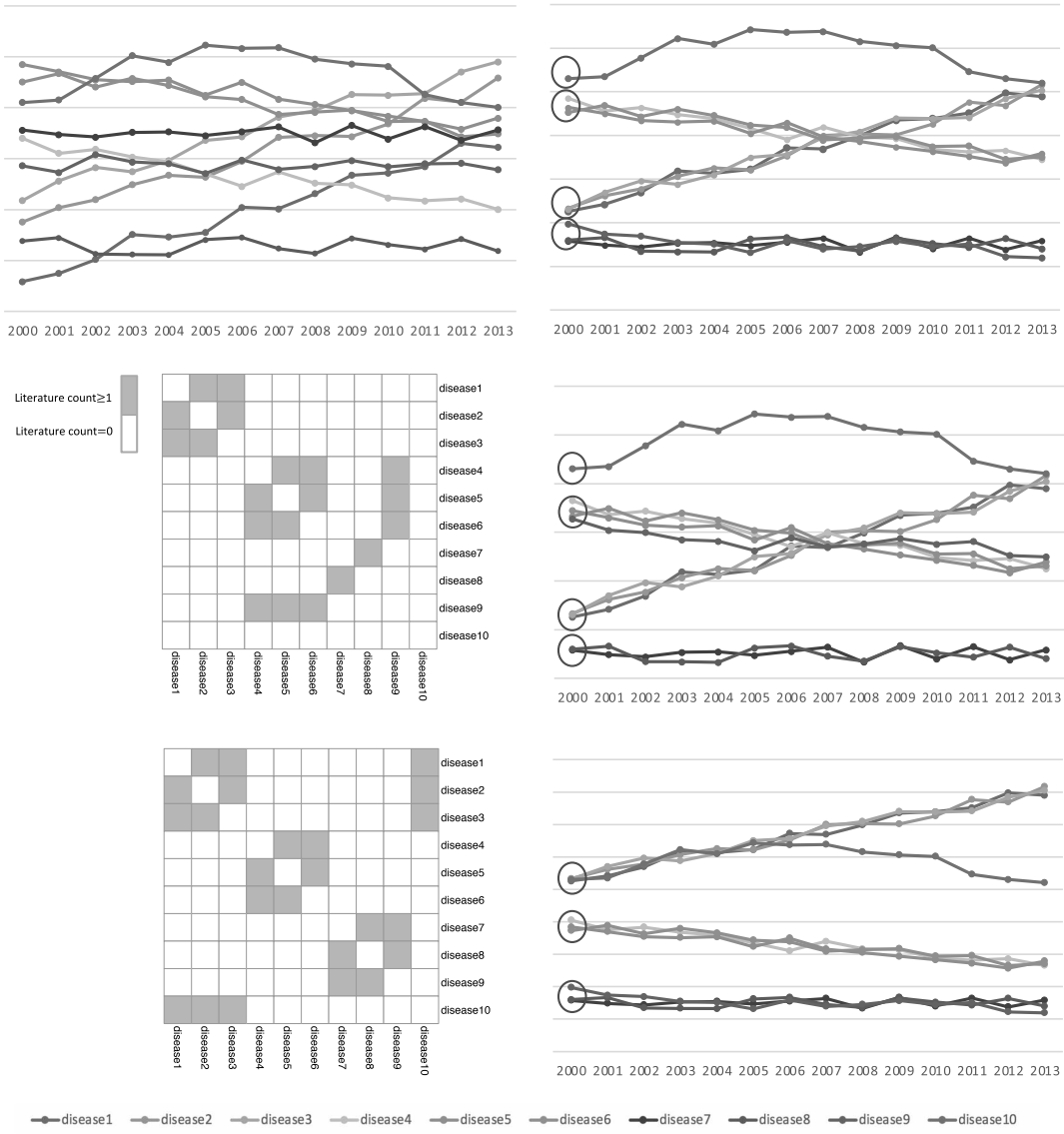

FIG. 1. *Upper-left*: *Original prevalence trends of* 10 *diseases*. *Upper-right*: *Clustering without incorporating existing information*. *Center-left and lower-left*: *Two existing information scenarios*. *Center-right and lower-right*: *Information-incorporated clusterings*.

by developing a novel strategy to incorporate existing information. Borrowing strength from outside information is not a new concept, however, limitedly pursued in functional clustering. In the middle and lower panels of Figure 1, we show two scenarios of existing information and the corresponding estimates when such information is incorporated, which suggests that incorporating information does have an impact on clustering. A foreseeable significant challenge is that information obtained from mining published studies can be partially correct or even incorrect. This is highly likely when it is not possible to accurately scrutinize each piece of information. For example, a study that mentions the interconnection between two diseases may derive that on a molecular basis. In this case the information is partially relevant. Consider another example where a study mentioning a disease pair actually suggests that they are not related. In this case this information will be included in Figure 2, however, is incorrect for the interconnection between the two diseases. To accommodate such scenar-

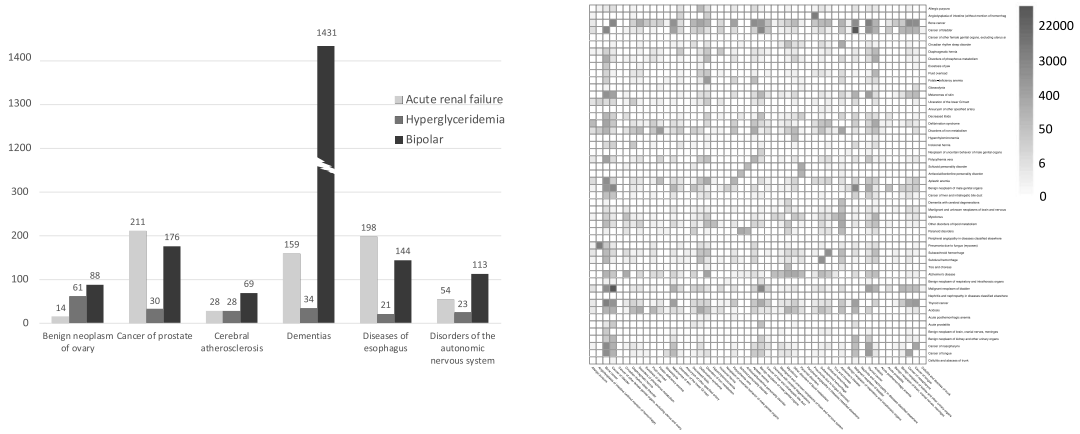

FIG. 2. *Existing information. Left*: *Barchart of literatures counts. Right*: *Heatmap of literature count for* 50 *selected diseases*.

ios, significant methodological developments are needed. *Another significant advancement from Jadhav et al. (2021) and some other studies is that rigorous theoretical development is conducted.* This can be nontrivial, as estimation needs to accommodate the partial correctness and incorrectness of existing information. Other methodological advancements, for example, the adoption of a different base penalty, are described below. Under the Bayesian paradigm, functional clustering using external information—such as publication results and published data—as prior has been developed (Biau et al. (2017), Isci et al. (2014), Ray and Mallick (2006)). The prior information and estimation strategies adopted in these studies are significantly different from those in this study.

Overall, the goal of this study is to conduct the clustering analysis of disease prevalence trends, with the assistance of information contained in published literature. This study advances from the existing ones in the following important aspects. First, the analysis scheme is significantly different and novel. More specifically, it differs from that limited to the prevalence of a single disease or a small number of diseases by conducting pan-disease analysis. It also differs from the HDN, pHDN, and pan-disease clinical treatment studies by analyzing prevalence. Second, the analysis technique significantly differs from the existing ones. The proposed approach is based on penalization fusion, which is relatively more recent and has notable advantages over many other functional clustering techniques. For example, it combines estimation and clustering and can conveniently determine the number of clusters. In principle, it can accommodate clusters as small as size one. Third, as described above, this study significantly advances from its closest competitor in Jadhav et al. (2021) both methodologically and statistically. Last but not least, it delivers a new way of extracting useful information from the NHIRD and other medical claims (record) databases.

2. Data. Two sources of data (Figure 3) are collectively analyzed. The first comes from the medical claims database and generates the prevalence values. The second comes from text mining and provides information on disease interconnections reported in the literature.

2.1. *NHI data*. In Taiwan basic health insurance coverage is provided by NHI, which was launched on March 1, 1995. By the end of 2014, almost 99.9% of the Taiwan population were covered. As NHI is also convenient and as uninsured and commercially insured healthcare is expensive, almost all hospital/clinic-based disease treatments have been going through NHI. In 2000, Taiwan established NHIRD, which contains detailed information on diagnosis, treatment, and outcome. We refer to the literature (Hsieh et al. (2019)) for more information

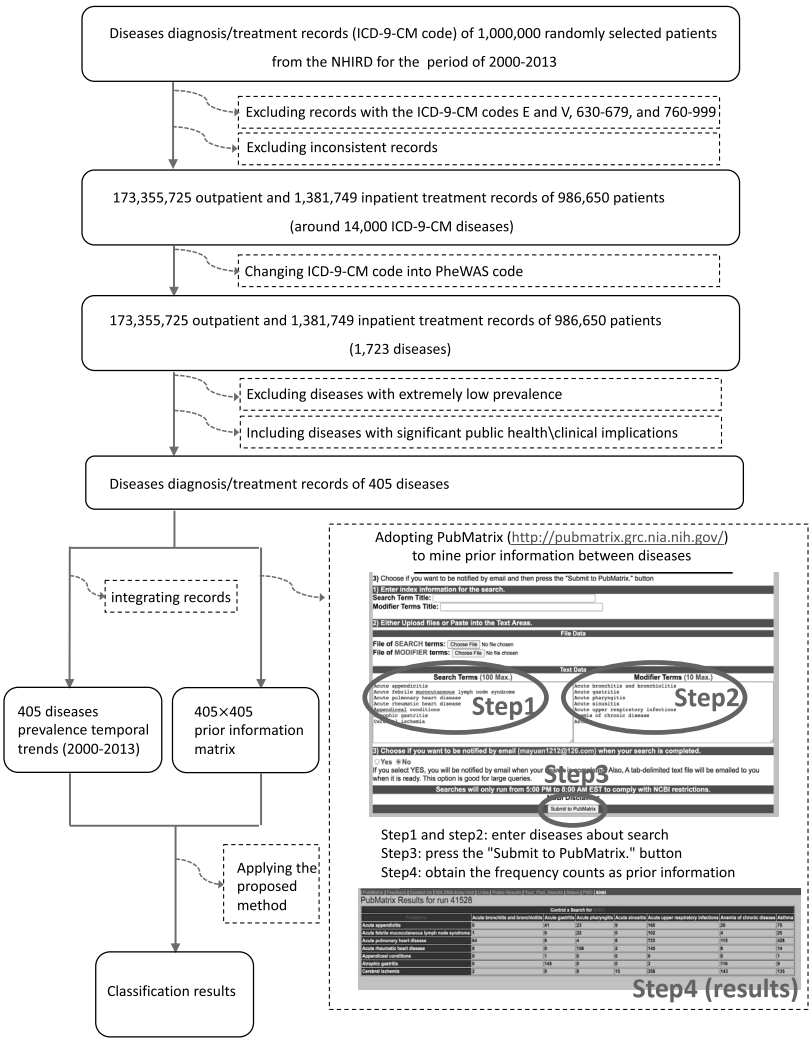

FIG. 3. *Flowchart of data preparation.*

on NHI and NHIRD. The NHI data has multiple unique characteristics: almost the whole population is covered; comprehensive information is available on all inpatient and outpatient treatment episodes; all data have been collected and stored under the same protocol, and extensive data processing has been conducted by NHIRD staff.

Data on one million randomly selected subjects is first extracted from NHIRD for the period of 2000–2013. Disease diagnosis and hence period prevalence information is collected from both outpatient and inpatient treatments, using the NHIRD CD (ambulatory care expenditure by visits) and DD (inpatient expenditure by admissions) files. For disease definition the ICD-9-CM code version 1992 is converted into the 2001 version. Records with the ICD-9-CM codes E and V (external causes of injury and supplemental classification), 630–679 (pregnancy, childbirth, and puerperium complications), and 760–999 (symptoms, signs, and ill-defined conditions) are excluded from analysis. The resulting dataset contains records on 986,650 patients, and disease period prevalence values are computed based on 173,355,725 outpatient and 1,381,749 inpatient treatment episodes. To avoid sparse data caused by too many diseases under ICD-9-CM, we apply the Phenome-Wide Association Study (PheWAS) vocabulary approach and group the 14,000 ICD-9-CM diseases into 1723 disease phecodes. We further select diseases based on the following considerations. We first identify diseases

that have high prevalence and/or high mortality, such as diseases of the circulatory system, certain cancers, diseases of the respiratory system, and others. We then also consider diseases that have high clinical significance (e.g., those with long inpatient length of stay, no effective treatment or clear causes), such as rare cancers, certain diseases of the blood and blood-forming organs, acquired coagulation factor deficiency, and others. Overall, 405 diseases are included in analysis (detailed information provided in the Supplementary Material (Ma et al. (2024))), and it is noted that this number is considerably larger than in many peer studies. These diseases have different types of trends (Figure 4).

2.2. *Information extraction*. There are multiple ways of defining and extracting existing information. The pan-disease perspective, variable of interest, and study population make our analysis unique. As such, it is unlikely the desired information can be obtained from a single or a few publications. We take a broad search strategy via text mining. Specifically, we adopt PubMatrix (pubmatrix.grc.nia.nih.gov), a web-based text mining tool tailored to PubMed, with the understanding that alternative tools such as VxInsight, MedMiner, and UALCAN may also be applicable. PubMatrix searches PubMed and returns the cooccurrence frequency of (i.e., number of publications that simultaneously contain) any pair of two keywords (e.g., "type 2 diabetes" and "melanoma"). We refer to Becker et al., 2003 and many publications that adopt this tool for details on how it functions. When considering all possible disease pairs, we can obtain results as exemplified in Figure 2. More detailed results are presented in the Supplementary Material S5. Here we note that this information extraction may be coarse. For example, different publications may call the same disease differently, and PubMatrix is not able to detect that. On the other hand, it is noted that the ICD-based disease phecode approach is standard. Our data analysis results suggest that a large number of publications can be identified using the PubMatrix-based text mining. It is possible to apply more advanced tools, which may be more complex, to refine the information; this may lead to improved estimation. As the proposed approach can accommodate partially correct information, this is not pursued. It is also noted that there are alternative ways of defining/extracting existing information on disease interconnections, for example, based on keywords of professional books, information related to clinical complications, cross-sectional epidemiological data, and so on. In Figure S2 (Supplementary Material), we also present the existing information based on the calculated disease prevalence correlation coefficients, where we see some similarities but also significant differences. The proposed information extraction can have notable advantages: it is based on a huge number of published studies and likely to be comprehensive and "less biased," and it can avoid the "using the same data twice" problem. We note that, in the literature, there is no optimal way of information extraction. As this is not our focus, we do not examine further.

3. Methods. Denote n as the number of diseases (sample size), T as the number of observations per disease, $y_i(t_j)$ as the prevalence value of disease i at the jth time point, and $\boldsymbol{Y_i} = (y_i(t_1), \ldots, y_i(t_T))^\top$ as the vector of observed prevalence values for disease i. Consider the model

$$y_i(t_j) = a_i + f_i(t_j) + \varepsilon_{ij}, \quad i = 1, \ldots, n, j = 1, \ldots, T, \tag{1}$$

where $t_j \in \mathcal{T}$, a_i's represent the average levels, $f_i(t)$'s are unknown smooth functions of t (with proper mean constraints for identifiability), and ε_{ij}'s are random errors. Let f_i^0 denote the true value of f_i. Assume that all curves can be classified into K clusters $\mathcal{G}_1, \ldots, \mathcal{G}_K$, and curves i and j belong to the same cluster if and only if $f_i^0 = f_j^0$. Let $\mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_K\}$. The goal is to simultaneously recover the structure of $\mathcal{G}$ and estimate the unknown f_i's. In what follows, we assume that normalization has been properly conducted, and a_i's are omitted.

3.1. *Penalized fusion*. We first briefly describe the penalized fusion analysis, as conducted in Jadhav et al. (2021), which is a building block of the proposed analysis. Consider the basis expansion

$$y_i(t_j) \approx \sum_{l=1}^{p} \beta_{il} x_l(t_j) + \varepsilon_{ij}, \quad i = 1, \ldots, n, \, j = 1, \ldots, T,$$

where $x_1(t), x_2(t), \ldots, x_p(t)$ are known basis functions and β_{il}'s are unknown regression coefficients. There have been extensive developments on choosing the form, number, and constraints of the basis functions (Schumaker (2007)), which are also applicable to this study. We adopt the B-spline basis in all our simulation and data analysis.

Denote $\boldsymbol{X} = \begin{bmatrix} x_1(t_1) & \cdots & x_p(t_1) \\ \vdots & \ddots & \vdots \\ x_1(t_T) & \cdots & x_p(t_T) \end{bmatrix}$, $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{ip})^\top$, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_n^\top)^\top$. Consider the loss function

$$Q(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X}) = \frac{1}{2T} \sum_{i=1}^{n} \|\boldsymbol{Y}_i - \boldsymbol{X}\boldsymbol{\beta}_i\|^2, \tag{2}$$

where $\boldsymbol{Y} = (\boldsymbol{Y}_1^\top, \ldots, \boldsymbol{Y}_n^\top)^\top$ and $\|\cdot\|$ is the ℓ_2 norm. We note that including covariance can in theory improve efficiency. However, our exploration suggests that, with penalization, the covariance estimates may be unsatisfactory, leading to overall inferior performance. As shown below, this loss function can lead to consistent clustering and estimation.

Consider the penalization fusion approach, which is abbreviated as "Fusion" in our numerical study. The penalized loss function is

$$Q_\lambda(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X}) = Q(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X}) + \sum_{i<j} p(\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|, \lambda), \tag{3}$$

where $p(\cdot, \lambda)$ is a concave penalty with a data-dependent tuning parameter $\lambda > 0$. In our numerical study, we adopt the minimax concave penalty (MCP, Zhang, 2010), defined by $p(t, \lambda) = \lambda \int_0^{|t|} (1 - x/(a\lambda))_+ \, dx$, and note that some other penalties are also applicable. Here $(x)_+ = xI(x > 0)$, and a is the regularization parameter. It is noted that Jadhav et al. (2021) adopts Lasso, which may have inferior properties compared to MCP. Denote the minimizer of (3) as $\hat{\boldsymbol{\beta}}^\lambda = (\hat{\boldsymbol{\beta}}_1^{\lambda\top}, \ldots, \hat{\boldsymbol{\beta}}_n^{\lambda\top})^\top$. Diseases i and j are concluded as in the same cluster if and only if $\hat{\boldsymbol{\beta}}_i^\lambda = \hat{\boldsymbol{\beta}}_j^\lambda$. We refer to Ma and Huang (2017) and follow-up studies for developments on penalized fusion.

3.2. *A new approach to incorporate existing information*. We propose a two-step approach to incorporate existing information on the interconnections among diseases:

In Step 1, consider the objective function

$$Q_\eta(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{W}) = Q(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X}) + \eta \sum_{i<j} w_{ij} \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|^2, \tag{4}$$

where $\eta > 0$ is a tuning parameter and $\boldsymbol{W} = (w_{ij})_{n\times n}$ is the weight matrix that describes existing information. In our numerical analysis, we set $w_{ij} = \log(1 + count_{ij})$, where $count_{ij}$ is the number of publications that simultaneously include diseases i and j. Denote the minimizer of (4) for disease i as $\hat{\boldsymbol{\beta}}_i^p$, compute the predicted value as $\hat{\boldsymbol{Y}}_i^p = \boldsymbol{X}\hat{\boldsymbol{\beta}}_i^p$, and denote $\hat{\boldsymbol{Y}}^p = (\hat{\boldsymbol{Y}}_1^{p\top}, \ldots, \hat{\boldsymbol{Y}}_n^{p\top})^\top$.

This is a weighted penalized estimation. In (4) if two diseases have more evidence of being interconnected, they are encouraged to have similar estimates. Similar strategies have been

developed in the literature, although under significantly different contexts. As the goal of this step is not to generate clustering, ridge-type penalization is imposed, which is computationally much simpler than penalized fusion. As can be seen from Figure 2, the distribution of the number of publications is quite skewed, which can be caused by research selection/publication bias, as opposed to the true amount of evidence. With this consideration and also to stabilize estimation, the logarithm transformation is taken.

In Step 2, we propose the objective function

$$Q_{\lambda,\tau}(\boldsymbol{\beta}; \boldsymbol{Y}, \hat{\boldsymbol{Y}}^p, \boldsymbol{X}) = (1-\tau)Q(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X}) + \tau Q(\boldsymbol{\beta}; \hat{\boldsymbol{Y}}^p, \boldsymbol{X}) + \sum_{i<j} p(\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|, \lambda), \tag{5}$$

where $0 \le \tau \le 1$ is a tuning parameter. λ has the same implication as in the standard penalized fusion. Simple derivation shows that the first two terms of (5) are equal to $Q(\boldsymbol{\beta}; \tilde{\boldsymbol{Y}}, \boldsymbol{X})$ plus a term that does not depend on $\boldsymbol{\beta}$, where $\tilde{\boldsymbol{Y}} = (1-\tau)\boldsymbol{Y} + \tau \hat{\boldsymbol{Y}}^p$. With a slight abuse of notation, denote the minimizer of (5) as $\hat{\boldsymbol{\beta}}$. The clustering structure can be fully obtained by examining $\hat{\boldsymbol{\beta}}$.

Objective function (5) has a very lucid interpretation. The loss balances between what is obtained from the data (the first term) and its counterpart if the existing information is credible (the second term). τ is introduced to balance between these two terms. Intuitively, if the information is of low quality, then with $\tau \to 0$, and the proposed analysis can reduce to that based on the observed data only. On the other hand, $\tau \to 1$ leads to the analysis heavily relying on the existing information. Assisted by this tuning, the proposed approach can flexibly accommodate a varying quality of existing information. For sparse linear regression, a related information-incorporation strategy has been developed by Jiang, He and Zhang (2016) from which this study advances by analyzing prevalence trends, conducting functional clustering, and extracting information in a different way.

In the proposed analysis, Jiang, He and Zhang (2016) and some others, the same data is analyzed in both steps. Our numerical and theoretical developments below as well as those in the published studies suggest that this is valid and sensible. We conjecture that it is also possible to split data and use one half for each step. When the number of subjects (for generating the prevalence data) is large enough, the two approaches are expected to have minimum differences.

3.3. *Statistical properties.* Consider model $y_i(t_j) = f_i(t_j) + \varepsilon_{ij}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, T$. Let $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iT})^\top$. Without loss of generality, assume $\mathcal{T} = [0, 1]$. With the rth order B-spline basis functions $x_1(t), \ldots, x_p(t)$, we have the approximation $f_i(t) \approx \sum_{l=1}^p \beta_{il} x_l(t)$. Here $p = m + r$, and m is the number of interior knots satisfying $0 = \kappa_0 < \kappa_1 < \cdots < \kappa_m < \kappa_{m+1} = 1$. By Corollary 6.21 of Schumaker (2007) and (C1)–(C3) listed below, there exists a spline approximation $\sum_{l=1}^p \beta_{il}^0 x_l(t)$ to f_i^0 such that $\sup_{i,t} |f_i^0(t) - \sum_{l=1}^p \beta_{il}^0 x_l(t)| = O(m^{-q})$, where $\boldsymbol{\beta}_i^0 = (\beta_{i1}^0, \ldots, \beta_{ip}^0)^\top$. Define $n_k = |\mathcal{G}_k|$, $\mathcal{G}_{\min} = \min_{1 \le k \le K} n_k$, where $|\mathcal{G}_k|$ is the size of $\mathcal{G}_k$. Let $\rho(t) = \lambda^{-1} p(t, \lambda)$ and $\mathcal{B}_{\mathcal{G}} = \{\boldsymbol{\beta} : \boldsymbol{\beta}_i = \boldsymbol{\beta}_j \text{ for any } i, j \in \mathcal{G}_k, 1 \le k \le K\}$. For a vector $\boldsymbol{a} = (a_1, \ldots, a_s)^\top \in \mathbb{R}^s$, let $\|\boldsymbol{a}\|_1 = \sum_{l=1}^s |a_l|$ and $\|\boldsymbol{a}\|_\infty = \max_{1 \le l \le s} |a_l|$. Assume the following conditions:

(C1) For each f_i^0 $(i = 1, \ldots, n)$, $f_i^0 \in C^q[0, 1]$ is a qth order continuously differentiable function defined on $[0, 1]$, and $r \ge q$.

(C2) Let $\delta = \max_{0 \le l \le m} (\kappa_{l+1} - \kappa_l)$. Assume that there exists a constant $M > 0$ such that $\frac{\delta}{\min_{0 \le l \le m} (\kappa_{l+1} - \kappa_1)} \le M$, $\max_{1 \le l \le m} |\kappa_{l+1} + \kappa_{l-1} - 2\kappa_l| = o(m^{-1})$, and $m = o(T)$.

(C3) For deterministic design points $t_i \in [0, 1]$, $i = 1, \ldots, T$, assume that there exists a distribution function G with a positive continuous design density $g(\cdot)$ such that $\sup_{t \in [0,1]} |G_n(t) - G(t)| = o(m^{-1})$, where $G_n(\cdot)$ is the empirical distribution of $t_1, \ldots, t_T$.

(C4) Assume that $\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ are independent, $E(\boldsymbol{\varepsilon}_i) = 0$, and $\max_{i,j} E(\varepsilon_{ij}^2) \le \sigma^2$. There exists $C_0 > 0$, such that $E[\exp\{(T^{-1}\boldsymbol{\varepsilon}_i^\top \boldsymbol{\varepsilon}_i)^{1/2}\}] \le C_0$ and $T^{-1}\sum_{t=1}^T \sum_{t'=1}^T |E[\varepsilon_{it}\varepsilon_{it'}]| \le C_0$ for $i = 1, \ldots, n$.

(C5) $\rho(t)$ is a symmetric function of t and is nondecreasing and concave in t for $t \in [0, \infty)$ with a continuous derivative $\rho'(t)$ on $(0, \infty)$. In addition, $\rho'(0+)$ is independent of λ. There exists a constant $0 < a < \infty$ such that $\rho(t)$ is a constant for all $t \ge a\lambda$.

(C1)–(C3) are standard assumptions for B-spline functions (Zhou, Shen and Wolfe (1998)). Condition (C4) gives the boundedness condition for the error terms (Chu, Li and Reimherr (2016)). Condition (C5) implies the choice of penalty functions and is common in the literature on high-dimensional variable selection (Fan and Lv (2011)). Both MCP and SCAD satisfy this condition.

First, consider the oracle estimator $\hat{\boldsymbol{\beta}}^{or}$ that incorporates the existing information,

$$\hat{\boldsymbol{\beta}}^{or} = \arg\min_{\boldsymbol{\beta} \in \mathcal{B}_\mathcal{G}} \frac{1}{2T} \sum_{i=1}^n \|\tilde{\boldsymbol{Y}}_i - \boldsymbol{X}\boldsymbol{\beta}_i\|^2,$$

where $\tilde{\boldsymbol{Y}}_i = (1-\tau)\boldsymbol{Y}_i + \tau \hat{\boldsymbol{Y}}_i^p$ with $\hat{\boldsymbol{Y}}_i^p = \boldsymbol{X}\hat{\boldsymbol{\beta}}_i^p$. Let $\sup_k \|\frac{1}{n_k}\sum_{i \in \mathcal{G}_k}(\hat{\boldsymbol{\beta}}_i^p - \boldsymbol{\beta}_i^0)\| = O_p(\psi_n)$. Note that this term reflects the reliability of the existing information.

THEOREM 3.1. *Suppose that Conditions* (C1)–(C4) *hold. Then we have*

$$\sup_i \|\hat{\boldsymbol{\beta}}_i^{or} - \boldsymbol{\beta}_i^0\| = O_p((1-\tau)\phi_n + \tau\psi_n),$$

where $\phi_n = m^{-q+1} + m\sqrt{K/(\mathcal{G}_{\min}T)}$.

Let $\boldsymbol{f}_i^0 = (f_i^0(t_1), \ldots, f_i^0(t_T))^\top$ and $b_n = \inf_{i \in \mathcal{G}_k, j \in \mathcal{G}_{k'}, k \ne k'} \frac{1}{\sqrt{T}}\|\boldsymbol{f}_i^0 - \boldsymbol{f}_j^0\|$. The following theorem shows that the oracle estimator $\hat{\boldsymbol{\beta}}^{or}$ is a strict local minimizer of the objective function with probability approaching one.

THEOREM 3.2. *Suppose that Conditions* (C1)–(C5) *hold. If* $\max(m^{-q}, m^{-1/2}\lambda) = o(b_n)$, $(1-\tau)\phi_n + \tau\psi_n = o(\lambda)$, $\mathcal{G}_{\min}^{-1}[(1-\tau)m^{-1/2}\log n + \tau m^{-1}\sup_i \|\hat{\boldsymbol{\beta}}_i^p - \boldsymbol{\beta}_i^0\|] = o_p(\lambda)$, *and* $\lambda = o(1)$, *where* ϕ_n *and* ψ_n *are given in Theorem* 3.1, *then there exits a strict local minimizer* $\hat{\boldsymbol{\beta}}$ *of the objective function* $Q_{\lambda,\tau}(\boldsymbol{\beta}; \boldsymbol{Y}, \hat{\boldsymbol{Y}}^p, \boldsymbol{X})$ *in* (5) *satisfying*

$$P(\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{or}) \to 1 \quad \text{as } n \to \infty.$$

Furthermore, $\sup_i \frac{1}{\sqrt{T}}\|\hat{\boldsymbol{f}}_i - \boldsymbol{f}_i^0\| = O_p((1-\tau)m^{1/2}(m^{-q} + \sqrt{K/(\mathcal{G}_{\min}T)}) + \tau(\psi_n + m^{-q}))$.

The above theorems show that, under mild conditions (including that on the estimate generated in Step 1 and hence the existing information), the proposed approach has clustering and estimation consistency. It can "automatically" determine the number of clusters. In line with Jiang, He and Zhang (2016), the assumed conditions and theoretical results are more complicated with functional data. Proof is provided in the Supplementary Material S1.

3.4. *Computation*. In the Supplementary Material S2, we develop an effective ADMM (alternating direction method of multipliers)-based algorithm. Information on tuning parameter selection is also presented.

3.5. *Simulation*. In the Supplementary Material S3, we conduct comprehensive simulation, evaluate performance of the proposed approach, and compare against multiple alternatives. It is observed that, when the existing information has moderate to high quality, the proposed approach significantly outperforms with higher clustering accuracy. When the existing information is of low quality, the alternative approach that fully trusts such information (i.e., Step 1 of the proposed analysis) may have inferior performance, while the proposed approach, with its flexibility, can correct for it to a large extent.

4. Data analysis. The NHIRD disease prevalence data and existing information extracted from PubMed, as described in Section 2, are analyzed. With selected $\tau = 0.40$, the proposed approach identifies 47 clusters with sizes at least 2. The nontrivial cluster sizes range from 2 to 28, with a median of 6. In addition, there are also eight diseases forming clusters with size 1. Detailed clustering information is provided in the Supplementary Material S5. For the nontrivial clusters, the unnormalized prevalence trends of their diseases are shown in Figure 4, where different colors correspond to different prevalence levels.

In general, we observe increasing trends in clusters 1–20. However, different clusters have different increasing patterns. For example, cluster 19 has increasing rates higher than the other clusters. This cluster includes atopic/contact dermatitis due to other or unspecified causes. This is a relatively common skin condition that affects a large number of children and adults in industrialized countries. It is estimated that about 19.5% of the general population in North America and Western Europe are affected. However, data has been relatively lacking for Taiwan. Environmental factors, which are often time-dependent, play an important role in the development of atopic/contact dermatitis, and aeroallergens are a trigger for exacerbations. The deterioration of air quality and other environmental factors in Taiwan has been well noted, which can explain the fast increase. Other diseases also included in cluster 19 include acute sinusitis, type 2 diabetes, reflux esophagitis, and mixed hyperlipidemia, whose increases over time have been reported for Taiwan in the literature. Similar sensible findings are also made with the other clusters with increasing trends. Representative examples include periodontitis in cluster 18, dysthymic disorder in cluster 16 as well as malignant and unknown neoplasms of brain and nervous system in cluster 3. Clusters 21–24 contain 50 diseases within general decreasing trends. Among these diseases neoplasm of uncertain behavior of breast has the sharpest decreasing trend, followed by disorders of esophageal motility and umbilical hernia. Some of the observed decreases can be explained by the discovery of advanced treatments, computerized devices, and improvements in healthcare services. For example, the decrease in the prevalence of viral hepatitis, particularly in industrialized nations, can be attributable to the effort in hepatitis vaccination, screening of blood products, screening and postexposure prophylaxis of healthcare workers, and increased availability of safe injection materials. A total of 96 diseases with relatively flat prevalence and small fluctuations are included in clusters 25–30. Their incidence may not be strongly influenced by time-dependent risk factors and prevention/diagnosis/treatment interventions. A total of 36 diseases in clusters 31–36 in general show reverse "V" shapes. Their prevalence values first rise to peaks and then decrease. The causes of such shapes have also been provided in the literature. Consider, for example, HPV. Invasive cervical cancer is one of the leading causes of cancer-related death among women. HPV vaccines were licensed in 2006 and then became widely available in many counties/regions including Taiwan. As such, a decrease in the prevalence of HPV infection and/or cervical cancer is sensible after the broad availability of vaccination. Diseases in clusters 37–38 have similar trends: their prevalence values first decrease and then stay relatively flat. In contrast, the prevalence values of diseases in clusters 39–40 first increase and then stay relatively flat. A total of 21 diseases with "irregular" trends are included in clusters 41-46. The "irregularity" can be caused by disease outbreaks, changes

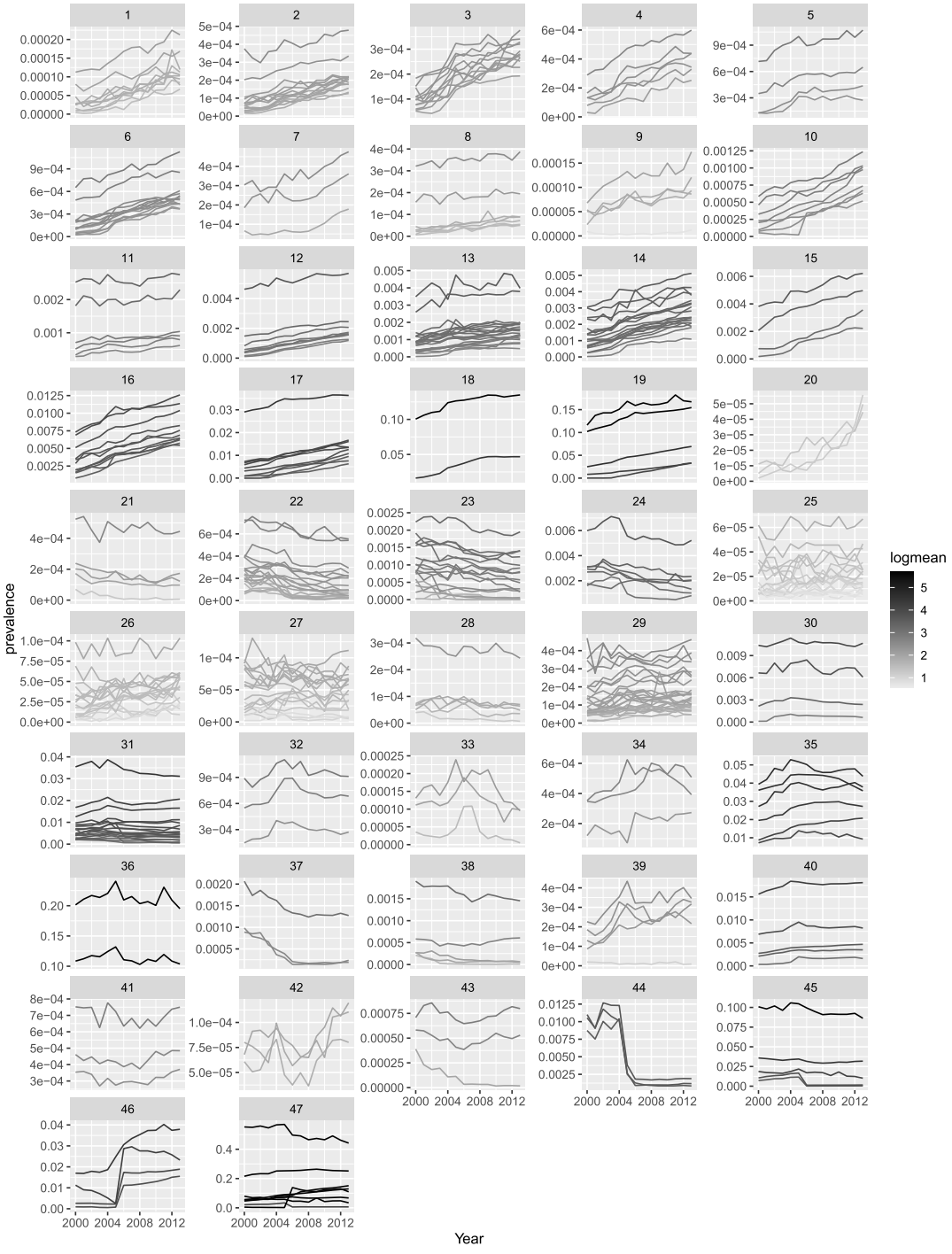

FIG. 4. *Clustering results using the proposed approach.*

in prevention and control measures, and interference with other related factors. The eight diseases forming their individual clusters are noninfectious gastroenteritis, gingivitis, essential hypertension, dental caries, acute gastritis, asthma with exacerbation, influenza, acute upper respiratory infections of multiple or unspecified sites. Most of these diseases are common, such as acute upper respiratory infections of multiple or unspecified sites, dental caries, and

gingivitis. Their high prevalence values may "amplify" variations, making them difficult to be clustered together with other diseases.

Although the above examination of individual prevalence trends for a large number of diseases is of interest, the key advancement of this study lies in the clustering analysis of diseases. A closer examination of the clustering results suggests their sensibility, with many of the interconnections reported in the literature (although in a very scattered manner). For example, it has been suggested that, because of shared genetic and other risk factors, the trends of the following diseases tend to be similar: HIV infection and hemangioma and lymphangioma (any site), which are coclustered in cluster 10 (Wiegand et al. (2008)), and renal failure NOS, thrombocytopenia, hypertensive heart and/or renal disease, and nephrotic syndrome without mention of glomerulonephritis, which are coclustered in cluster 13 (Kressel et al. (1981)). Another sensible finding is that disorders of function of stomach, gastritis and duodenitis, and peptic ulcer (excluding esophageal) are coclustered in cluster 44. It has been suggested by Sipponen and Hyvärinen (1993) that the pathogenesis of peptic ulcer and gastric cancer is closely associated with H. pylori gastritis and its subsequent atrophic sequelae (atrophic gastritis). Cellulitis is a spreading bacterial infection of the skin and tissues immediately beneath the skin (Gabillot-Carré and Roujeau (2007)). With this cause other local infections of skin and subcutaneous tissue and cellulitis and abscess of arm/hand are clustered together in cluster 45. Acute bronchitis and bronchiolitis and acute pharyngitis are the only two diseases in cluster 36, and they have very similar prevalence trend patterns. This is because acute bronchitis and bronchiolitis often develops from other upper respiratory tract infections, such as acute pharyngitis. Beyond those with strong support from the literature, there are also new disease coclusterings that have not been suggested in the literature. For example, it is found that schizoid personality disorder and dissociative disorder share similar patterns and are coclustered in cluster 8, although published studies suggest that their interconnections are inconclusive (Modestin, Hermann and Endrass (2007)). Another example is that cancer of lip and allergic purpura, which are coclustered in cluster 1, have similar prevalence patterns but do not have support from published literature. It is noted that the prevalence values are computed based on a large number of individuals. Their credibility is expected to be high, and as such, the observed similarity in prevalence patterns is expected to be true. The above new findings on coclusterings and those alike suggest new directions for identifying interconnections underneath diseases.

Data is also analyzed using the following alternatives, which are also considered in simulation and described in more detail in the Supplementary Material S3. [OLS] Each prevalence curve is first estimated separately. Then diseases i and j are clustered together if $\|\hat{\boldsymbol{\beta}}_i - \hat{\boldsymbol{\beta}}_j\| \leq \kappa$. Here κ is determined in a similar way as for the proposed approach (Supplementary Material S2). [kmeans] This approach first fits each prevalence curve separately. Then the vectors of regression coefficients are clustered using the kmeans approach. [distK] This approach first computes distance correlation between the observed prevalence values and then use the Kmeans method to generate clusters. It is implemented using the R functions *dcor* and *kmeans*. [funFEM] This method is based on mixture modeling and takes functions as input. As such, we first implement a smoothing method to obtain functions passing through the observed discretized points. This method is implemented using the R package *funFEM*. [FClust] The FClust approach is implemented using the R package *fdapace* and conducts functional clustering and identification of data substructures for longitudinal and other functional data. [funHDDC] The funHDDC method is implemented using the R package *funHDDC* and conducts model-based clustering and identification of functional subspaces. [waveclust] This method is based on a wavelet decomposition of signal and a mixture model that integrates random effects and implemented with the R package *curvclust*. [fitfclust] This is a functional clustering method with special attention to sparsely sampled data and available

through the R package *funcy*. [Fusion] This is the approach described in Section 3.1. [Prior] This approach generates estimates as described in Step 1 of the proposed approach and then conducts clustering in the same way as the OLS approach.

The clustering results using the alternatives are presented in Figures S3–S12 (Supplementary Material S4) and the Supplementary Material S5. The resulted numbers of clusters are 22 (OLS), 10 (kmeans), 10 (distK), two (funFEM), 10 (Fclust), three (funHDDC), 10 (waveclust), 10 (fitclust), 35 (Fusion), and 17 (prior). In Table S5 (Supplementary Material S4), we present the discrepancy (which is the normalized "clustering error") between any two methods and observe small to large discrepancy values. Here we note that, with a large number of disease pairs, a small discrepancy value can correspond to notable differences in clustering, which can be partly reflected in the number of clusters and numbers of diseases in clusters. With respect to the proposed approach, the Fusion approach leads to the most similar findings (with a discrepancy value of 0.06), and the funHDDC approach leads to the most discrepant findings (with a discrepancy value of 0.9). This is reasonable as the analysis frameworks of the proposed and Fusion approaches are closest, while funHDDC has a highly different framework.

We take a closer look at Fusion. The moderate τ value and relatively small discrepancy value seem to suggest a small impact of the existing information. However, a closer look suggests that Fusion and the proposed approach have significant differences in clustering structures (e.g., number of clusters, number of diseases in individual clusters, and disease memberships). A representative example is in Figure S13 (Supplementary Material S4). Specifically, under the Fusion approach, eight diseases are clustered together in cluster 7. Incorporating the existing information, they belong to four different clusters—along with other diseases—under the proposed approach (detailed information available from the authors). Another example is that the proposed approach clusters malignant and unknown neoplasms of brain and nervous system with cancer of brain in cluster 3. However, under Fusion, they belong to different clusters. One more example of a similar kind is reticulosarcoma and benign neoplasm of adrenal gland, which are coclustered in cluster 29. This interconnection can be partly explained by p53 tumor-suppressor gene and Li-Fraumeni syndrome. Germline mutations in p53 have been identified in families with the Li-Fraumeni syndrome, a rare familial cancer syndrome characterized by an unusually high incidence of multiple cancers such as sarcomas, adrenocortical carcinomas, and other diverse neoplasms. Families with Li-Fraumeni syndrome have been described as including a proband with a sarcoma diagnosed early in life (Malkin et al. (1992)).

To gain further insights, we also conduct evaluation. First, we use the first 13 observations of all diseases for clustering analysis and estimation. Then, based on the models, the last observation of each disease is predicted. The mean squared error values are 1.29×10^{-5} (OLS), 8.32×10^{-6} (Fusion), 1.12×10^{-5} (Prior), and 7.28×10^{-6} (proposed), respectively. Here we note that the other alternatives do not generate explicit regression models and hence cannot be directly used for prediction. This evaluation demonstrates that, as it generates models, the proposed analysis can also be used for prediction. As it is not the focus of this analysis, we do not further pursue this aspect. In addition, we conduct an evaluation of the stability of clustering. Specifically, we randomly select 2/3 of the diseases and apply the proposed and alternative approaches. We then compare the clustering with the randomly sampled data against that using the whole data and compute the discrepancy value. Note that this calculation is limited to diseases that are selected. With 500 resamplings the resulted stability measure (1-discrepancy) values are 0.966 (OLS), 0.973 (Fusion), 0.962 (Prior), and 0.977 (proposed), respectively. Overall, the proposed approach has competitive prediction and stability performance.

5. Discussion. In this study we have conducted the functional clustering analysis of disease prevalence trends, taking advantage of the uniquely valuable NHI data. Similar analysis schemes have been limitedly pursued in the literature, and as discussed above, the findings can have high practical value. Methodologically, this study has significantly advanced from Jadhav et al. (2021) and other functional clustering analyses by flexibly incorporating existing information obtained from mining a large number of PubMed publications. The proposed approach is intuitive and will be directly applicable to some other types of existing information. Also different from Jadhav et al. (2021), the MCP penalization, which has been shown as more effective than Lasso, is adopted. Significant theoretical development has been conducted, which is highly nontrivial with the complex data structure, penalized fusion estimation, and existing information that is not guaranteed to be accurate. Simulation shows that when existing information has reasonable quality, the proposed approach has superior performance. When existing information is of very low quality, which is highly unlikely in practice, the proposed approach, with its great flexibility, can still have competitive performance.

In the analysis of NHI data, clustering results different from the alternatives have been generated. It is also found that incorporating the existing information has a moderate impact on the results. For a few cases, for example, the one related to HPV, we have identified solid evidence from the literature to support our findings. Here we note that although for some diseases published studies have suggested their similarity in prevalence trends and underneath interconnections, this study is among the first to systematically do so at the pan-disease level. It is also noted that a drawback of pan-disease analysis is that, with a huge number of disease pairs, it is impossible to examine the underlying causes one by one. Rather our findings can serve as the ground stone for researchers interested in individual diseases. Although with many advantages, the NHI data also has limitations. In particular, the covered population is dominatingly Chinese. In addition, with data access limitation, more recent data is not available. With the dependence of disease prevalence on population and time, the broader applicability of our findings may warrant additional scrutinization.

Software and data. The R program implementing the proposed method is available at www.github.com/shuanggema. The data that support the findings in this paper are obtained from the National Health Insurance Research Database at https://nhird.nhri.org.tw/en/.

Acknowledgments. The authors thank the Editor and reviewers for their careful review and insightful comments, which have led to a major improvement of this article.

Funding. This study was supported by China Postdoctoral Science Foundation (2022M 720328), CSIAM Research Project for Young Women in Applied Mathematics, Beijing Postdoctoral Research Foundation, National Natural Science Foundation of China (11971404, 11701561), 111 Project (B13028), National Statistical Science Research Project (2019LZ22), Fund for building world-class universities (disciplines) of Renmin University of China, NSF (1916251, 2209685), and a Yale Cancer Center Pilot Award.

SUPPLEMENTARY MATERIAL

Supplementary Material S1–S5 (DOI: 10.1214/23-AOAS1821SUPPA; .pdf). The supplement provides more details on the proofs of the theoretical results (Supplementary Material S1), details on computation (Supplementary Material S2), simulation design and results (Supplementary Material S3), and additional data analysis results (Supplementary Material S4 and S5).

Source code (DOI: 10.1214/23-AOAS1821SUPPB; .zip). R programs for implementing the proposed method.

REFERENCES

BECKER, K. G., HOSACK, D. A., DENNIS, G., LEMPICKI, R. A., BRIGHT, T. J., CHEADLE, C. and ENGEL, J. (2003). PubMatrix: A tool for multiplex literature mining. *BMC Bioinform*. **4** 1–6.

BIAU, D. J., BOULEZAZ, S., CASABIANCA, L., HAMADOUCHE, M., ANRACT, P. and CHEVRET, S. (2017). Using Bayesian statistics to estimate the likelihood a new trial will demonstrate the efficacy of a new treatment. *BMC Med. Res. Methodol*. **17** 1–10.

CANTO, J. G., SHLIPAK, M. G., ROGERS, W. J., MALMGREN, J. A., FREDERICK, P. D., LAMBREW, C. T., ORNATO, J. P., BARRON, H. V. and KIEFE, C. I. (2000). Prevalence, clinical characteristics, and mortality among patients with myocardial infarction presenting without chest pain. *JAMA* **283** 3223–3229. https://doi.org/10.1001/jama.283.24.3223

CHEN, L. L., BLUMM, N., CHRISTAKIS, N. A., BARABÁSI, A.-L. and DEISBOECK, T. S. (2009). Cancer metastasis networks and the prediction of progression patterns. *Br. J. Cancer* **101** 749–758. https://doi.org/10.1038/sj.bjc.6605214

CHU, W., LI, R. and REIMHERR, M. (2016). Feature screening for time-varying coefficient models with ultrahigh-dimensional longitudinal data. *Ann. Appl. Stat*. **10** 596–617. MR3528353 https://doi.org/10.1214/16-AOAS912

FAN, J. and LV, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Inf. Theory* **57** 5467–5484. MR2849368 https://doi.org/10.1109/TIT.2011.2158486

GABILLOT-CARRÉ, M. and ROUJEAU, J.-C. (2007). Acute bacterial skin infections and cellulitis. *Curr. Opin. Infect. Dis*. **20** 118–123. https://doi.org/10.1097/QCO.0b013e32805dfb2d

GOH, K.-I., CUSICK, M. E., VALLE, D., CHILDS, B., VIDAL, M. and BARABÁSI, A.-L. (2007). The human disease network. *Proc. Natl. Acad. Sci. USA* **104** 8685–8690.

HSIEH, C.-Y., SU, C.-C., SHAO, S.-C., SUNG, S.-F., LIN, S.-J., YANG, Y.-H. K. and LAI, E. C.-C. (2019). Taiwan's national health insurance research database: Past and future. *Clin. Epidemiol*. **11** 349.

ISCI, S., DOGAN, H., OZTURK, C. and OTU, H. H. (2014). Bayesian network prior: Network analysis of biological data using external knowledge. *Bioinformatics* **30** 860–867. https://doi.org/10.1093/bioinformatics/btt643

JACQUES, J. and PREDA, C. (2014). Functional data clustering: A survey. *Adv. Data Anal. Classif.* **8** 231–255. MR3253859 https://doi.org/10.1007/s11634-013-0158-y

JADHAV, S., MA, C., JIANG, Y., SHIA, B.-C. and MA, S. (2021). Pan-disease clustering analysis of the trend of period prevalence. *Ann. Appl. Stat*. **15** 1945–1958. MR4355083 https://doi.org/10.1214/21-aoas1470

JIANG, Y., HE, Y. and ZHANG, H. (2016). Variable selection with prior information for generalized linear models via the prior LASSO method. *J. Amer. Statist. Assoc*. **111** 355–376. MR3494665 https://doi.org/10.1080/01621459.2015.1008363

JOFFRES, M., FALASCHETTI, E., GILLESPIE, C., ROBITAILLE, C., LOUSTALOT, F., POULTER, N., MCALISTER, F. A., JOHANSEN, H., BACLIC, O. et al. (2013). Hypertension prevalence, awareness, treatment and control in national surveys from England, the USA and Canada, and correlation with stroke and ischaemic heart disease mortality: A cross-sectional study. *BMJ Open* **3** e003423. https://doi.org/10.1136/bmjopen-2013-003423

KRESSEL, B. R., RYAN, K. P., DUONG, A. T., BERENBERG, J. and SCHEIN, P. S. (1981). Microangiopathic hemolytic anemia, thrombocytopenia, and renal failure in patients treated for adenocarcinoma. *Cancer* **48** 1738–1745. https://doi.org/10.1002/1097-0142(19811015)48:8<1738::aid-cncr2820480808>3.0.co;2-e

LAI, Y.-H. (2015). Network analysis of comorbidities: Case study of HIV/AIDS in Taiwan. In *International Conference on Multidisciplinary Social Networks Research* 174–186. Springer, Berlin.

LUCHSINGER, J. A., REITZ, C., PATEL, B., TANG, M.-X., MANLY, J. J. and MAYEUX, R. (2007). Relation of diabetes to mild cognitive impairment. *Arch. Neurol*. **64** 570–575.

MA, C., LI, Y., SHIA, B. and MA, S. (2020). Human disease cost network analysis. *Stat. Med*. **39** 1237–1249. MR4098487 https://doi.org/10.1002/sim.8472

MA, C., LIN, C., XUE, Y., ZHANG, S., ZHANG, Q. and MA, S. (2024). Supplement to "Information-incorporated clustering analysis of disease prevalence trends." https://doi.org/10.1214/23-AOAS1821SUPPA, https://doi.org/10.1214/23-AOAS1821SUPPB

MA, S. and HUANG, J. (2017). A concave pairwise fusion approach to subgroup analysis. *J. Amer. Statist. Assoc*. **112** 410–423. MR3646581 https://doi.org/10.1080/01621459.2016.1148039

MALKIN, D., JOLLY, K. W., BARBIER, N., LOOK, A. T., FRIEND, S. H., GEBHARDT, M. C., ANDERSEN, T. I., BØRRESEN, A.-L., LI, F. P. et al. (1992). Germline mutations of the p53 tumor-suppressor gene in children and young adults with second malignant neoplasms. *N. Engl. J. Med*. **326** 1309–1315.

MODESTIN, J., HERMANN, S. and ENDRASS, J. (2007). Schizoidia in schizophrenia spectrum and personality disorders: Role of dissociation. *Psychiatry Res*. **153** 111–118. https://doi.org/10.1016/j.psychres.2006.03.003

RAY, S. and MALLICK, B. (2006). Functional clustering by Bayesian wavelet methods. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **68** 305–332. MR2188987 https://doi.org/10.1111/j.1467-9868.2006.00545.x

ROMANOWSKI, M. D., PAROLIN, M. B., FREITAS, A. C. T., PIAZZA, M. J., BASSO, J. and URBANETZ, A. A. (2015). Prevalence of non-alcoholic fatty liver disease in women with polycystic ovary syndrome and its correlation with metabolic syndrome. *Arq. Gastroenterol.* **52** 117–123. https://doi.org/10.1590/S0004-28032015000200008

SCHUMAKER, L. L. (2007). *Spline Functions*: *Basic Theory*, 3rd ed. *Cambridge Mathematical Library*. Cambridge Univ. Press, Cambridge. MR2348176 https://doi.org/10.1017/CBO9780511618994

SIPPONEN, P. and HYVÄRINEN, H. (1993). Role of Helicobacter pylori in the pathogenesis of gastritis, peptic ulcer and gastric cancer. *Scand. J. Gastroenterol.* **28** 3–6.

TSAI, C.-P., HU, C. and LEE, C. T.-C. (2019). Finding diseases associated with amyotrophic lateral sclerosis: A total population-based case–control study. *Amyotroph. Lateral Scler. Frontotemporal Degeneration* **20** 82–89.

WIEGAND, S., EIVAZI, B., BARTH, P. J., VON RAUTENFELD, D. B., FOLZ, B. J., MANDIC, R. and WERNER, J. A. (2008). Pathogenesis of lymphangiomas. *Virchows Arch.* **453** 1–8. https://doi.org/10.1007/s00428-008-0611-z

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701 https://doi.org/10.1214/09-AOS729

ZHOU, S., SHEN, X. and WOLFE, D. A. (1998). Local asymptotics for regression splines and confidence regions. *Ann. Statist.* **26** 1760–1782. MR1673277 https://doi.org/10.1214/aos/1024691356

ZHOU, X., LEI, L., LIU, J., HALU, A., ZHANG, Y., LI, B., GUO, Z., LIU, G., SUN, C. et al. (2018). A systems approach to refine disease taxonomy by integrating phenotypic and molecular networks. *eBioMedicine* **31** 79–91.

ZHOU, X., MENCHE, J., BARABÁSI, A.-L. and SHARMA, A. (2014). Human symptoms–disease network. *Nat. Commun.* **5** 1–10.