Completing Missing Prevalence Rates for Multiple Chronic Diseases by Jointly Leveraging Both Intra- and Inter-Disease Population Health Data Correlations

Yujie Feng
Peking University
China
yj.feng@pku.edu.cn

Jiangtao Wang*
Coventry University
United Kingdom
jiangtao.wang@coventry.ac.uk

Yasha Wang*
Peking University
China
wangys@sei.pku.edu.cn

Sumi Helal
University of Florida
United States
helal@cise.ufl.edu

ABSTRACT

Population health data are becoming more and more publicly available on the Internet than ever before. Such datasets offer a great potential for enabling a better understanding of the health of populations, and inform health professionals and policy makers for better resource planning, disease management and prevention across different regions. However, due to the laborious and high-cost nature of collecting such public health data, it is a common place to find many missing entries on these datasets, which challenges the utility of the data and hinders reliable analysis and understanding. To tackle this problem, this paper proposes a deep-learning-based approach, called Compressive Population Health (CPH), to infer and recover (to complete) the missing prevalence rate entries of multiple chronic diseases. The key insight of CPH relies on the combined exploitation of both intra-disease and inter-disease correlation opportunities. Specifically, we first propose a Convolutional Neural Network (CNN) based approach to extract and model both of these two types of correlations, and then adopt a Generative Adversarial Network (GAN) based prevalence inference model to jointly fuse them to facility the prevalence rates data recovery of missing entries. We extensively evaluate the inference model based on real-world public health datasets publicly available on the Web. Results show that our inference method outperforms other baseline methods in various settings and with a significantly improved accuracy (from 14.8% to 9.1%).

CCS CONCEPTS

• **Applied computing** → **Data recovery**; *Health care information systems.*

KEYWORDS

population health, missing data recovery, generative adversarial network

ACM Reference Format:
Yujie Feng, Jiangtao Wang, Yasha Wang, and Sumi Helal. 2021. Completing Missing Prevalence Rates for Multiple Chronic Diseases by Jointly Leveraging Both Intra- and Inter-Disease Population Health Data Correlations. In *Proceedings of the Web Conference 2021 (WWW '21), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3442381.3449811

1 INTRODUCTION

Due to unhealthy lifestyles and a fast-growing ageing population, many chronic and malignant diseases (e.g., heart diseases, diabetes, and cancer) are becoming increasingly prevalent in our society. Understanding the changes in population health patterns and trends is crucial for the monitoring, resource planning, and evaluation of health programs and policies. To achieve these goals, population health surveillance, which is a typical institutionalized sensing of information about the health status of a population, has become a core function for a nation's public health system. The World Health Organization (WHO) lists population health monitoring as the first of ten essential public health operations [29]. Profiling spatially fine-grained prevalence rate (morbidity rate) of multiple chronic diseases is a critical task in population health surveillance [24], which helps public health decision-makers, health planning administrators, pharmaceutical manufacturers, and clinicians, to effectively treat diseases, allocate medical resources, and manage population health.

The increasing availability of publicly available datasets for disease prevalence rates provides new opportunities for healthcare researchers/professionals to better study and understand public health and wellbeing from multiple perspectives (e.g., the study of health inequality by spatial epidemiologist). However, the road to constructing and publicly availing these population health datasets is challenging. To appreciate the challenges we briefly describe the processes by which data is collected. There are commonly two ways for healthcare authorities to perform prevalence profiling, namely clinical-record integration [20] and residents survey [22]. For healthcare authorities that adopt clinical-record integration,

*Corresponding author.

WWW '21, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8312-7/21/04.
https://doi.org/10.1145/3442381.3449811

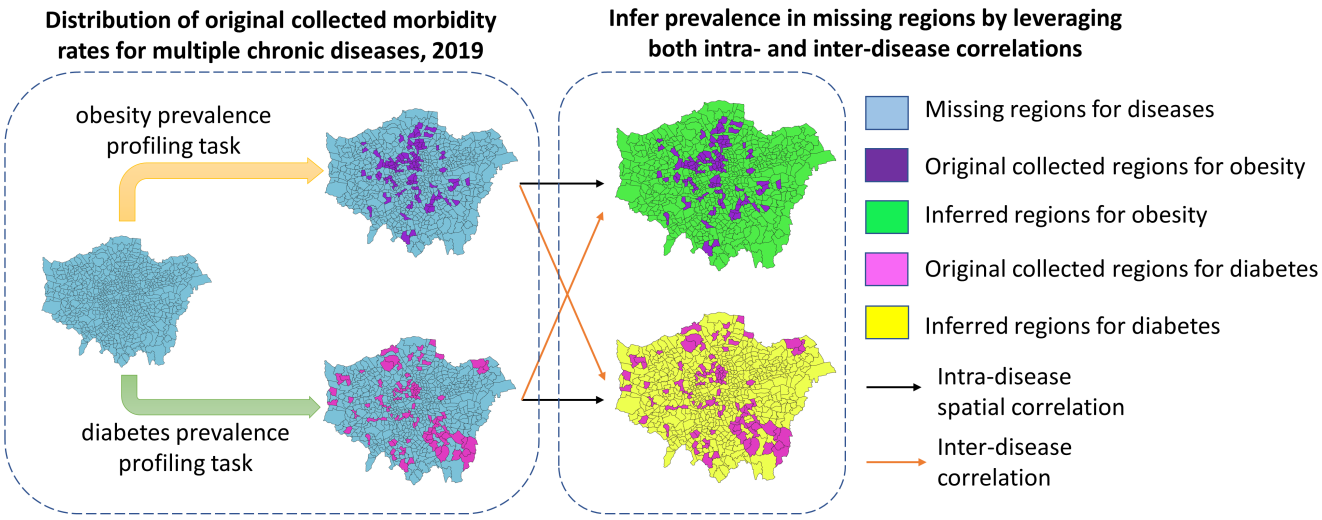


Figure 1: Basic vision of compressive population health (CPH)

they need to integrate data sourced from heterogeneous information systems belonging to multiple medical institutions, to get an overview of morbidity rates. However, such data integration task is non-trivial due to the many reasons. First, data access while preserving patient privacy comes at a non-trivial cost and overhead. For example, to know the population health statistics in a city, we need to access individual-level sensitive health data. Thus, health organizations must take on a lot of laborious work to ensure data anonymity (and hence patient privacy) before feeding into data integration. Second, the data structure and database design might be different for heterogeneous systems from multiple clinics or hospitals, thus increasing the difficulty and cost of data linkage. For healthcare authorities that adopt residents surveying, they need to recruit a representative group of residents and collect data via interviews or self-managed questionnaires. In order to minimize the bias, the number of samples should be large enough with appropriate demographic distributions. Therefore, this process is time-consuming and incurs high cost including labor cost for survey administrators and incentive payments for survey participants.

As a result of the aforementioned difficulties in health data integration and prevalence profiling, the publicly released population health datasets are often found to be incomplete, where the prevalence rates for some regions or for some types of diseases are missing. The data incompleteness significantly lower the quality and power of the released data, which hinders timely data analysis and reliable knowledge generation by epidemiologists or public health authorities or researchers. Incompleteness also exacerbates the uncertainty in verifying or studying health equity as noted in [25].

To bridge this gap, this paper introduces a novel approach, named Compressive Population Health (CPH), for completing the missing entries of prevalence rates of multiple chronic diseases thus re-constructing a full, reliable and timely population health monitoring picture for the current year. The key insight of CPH is that the missing entries of prevalence rates of the current year could be "recovered" by leveraging both intra-disease spatial correlations as well as inter-disease correlations (see Figure 1). For intra-disease spatial correlations, a number of studies have highlighted the role of neighborhood effects on health, that is, nearby regions are more similar in the prevalence of certain diseases than distant ones [2]. This is because nearby regions share common environmental, socioeconomic, and demographic features. For inter-disease correlations, multi-morbidity [8], commonly defined as the co-presence of two or more diseases, demonstrates that statistics for different types of diseases may also correlate with each other. For example, regions with higher obesity rate are more likely to have higher rates of ischemic heart disease and cancers [8]. We assume that both of these two types of correlations will be learned from both historical prevalence data and the known data entries of the current target year.

Although the above two types of data correlations have been demonstrated in the area of epidemiology, there are still technical challenges to realize the vision of CPH. (1) Challenge A: how to extract and model both intra- and inter-disease data correlations based on incomplete training data. (2) Challenge B: how to jointly represent and fuse these correlations to build an accurate prevalence inference model. To the best of our knowledge, these challenges have not been touched in the area of epidemiology. In the computer science community, although completing missing data entries with spatial correlations has been studied in other domains such as environmental and traffic monitoring [10][33], they only focused on a single task so that cannot effectively incorporate inter-disease correlations into the data recovery tasks of multiple diseases in our focused scenario.

The contributions of this paper are summarized as follows:

1) This paper introduce a novel approach, named Compressive Population Health (CPH), for completing the missing prevalence rate of multiple chronic diseases of the current year so that a timely and full picture of public health surveillance can be re-constructed. The missing data of the current year is recovered by leveraging

both intra-disease spatial correlations and inter-disease correlations, which are learned based on prevalence data from both historical years and known entries of the current year.

2) We propose a deep-learning-based prevalence inference model to jointly utilize both intra- and inter-disease data correlations. Specifically, the inference model consists of two main components: the Convolutional Neural Network (CNN) based method is designed to extract and represent both the intra-disease spatial correlations and inter-disease correlations (Responds to Challenge A), while a Generative Adversarial Network (GAN) based model is utilized to make inference by combining these two types of correlations with high accuracy (Responds to Challenge B).

3) We extensively evaluate the inference model based on real-world public health datasets publicly available on the Web containing three types of correlated diseases over ten years, and the results show that our data recovery method outperforms other baselines on various settings with an improved accuracy of 14.8% to 9.1%.

The rest of the paper is organized as follows: Section 2 formally defines the problem. Section 3 introduces the missing data inference methodology for CPH. Section 4 evaluates the effectiveness of CPH based on real-world population health datasets. Section 5 reviews related works. In Section 6, we discuss limitations of this work and future work directions. Section 7 concludes this study. The code to reproduce the experiment is available at [1].

2 PROBLEM FORMULATION

2.1 Motivating Example

Suppose that the public health authority of a certain region (e.g., Greater London) has collected and integrated the prevalence rate of multiple diseases such as obesity, hypertension, and diabetes, across 500+ grids (e.g., wards) for the year 2019, and now intends to share them on the Web as part of the government data open plan so that epidemiologist can investigate the health inequality of this year in a timely manner. Due to the difficulties in data collection and integration (either survey-based approach or clinic-visit-based data integration approach described in Section 1), they found that there are some missing data entries for certain grids or diseases. Therefore, they first need to use CPH to complete the missing entries before sharing them on the Web. Given the recent years' historical prevalence statistics provided by the public health authority before 2019 (e.g., from 2015 to 2018) and available data in 2019, CPH is able to extract and model both inter-disease and intra-disease data correlations, and then re-construct the full health surveillance map across the different grids for 2019.

2.2 Formal Problem Definition

We define a dataset $\mathcal{X}$ containing multiple chronic diseases, where $\mathcal{X} = \{X^1, X^2, \ldots, X^S\}$, and where S denotes the chronic diseases, and X denotes disease matrix. For each disease matrix X, where a row stands for a region and a column denotes a timestamp. An entry X_{it} refers to the reading of i^{th} region at t^{th} timestamp. Under the basic idea of CPH, the disease datasets are incomplete, e.g. some regions or some type of diseases are missing. So, in a disease matrix X, we call the missing regions as missing entries and the regions with available (original) values as observed entries. Then we define a binary mask M in the following way:

$$M_{it} = \begin{cases} 1, & \text{if } X_{it} \text{ is observed} \\ 0, & \text{if } X_{it} \text{ is missing} \end{cases} \tag{1}$$

so that M indicates which entries of X are observed.

We define two parameters: target year and sampling proportion $\mathcal{R}$, where target year is the year for which we need to perform the data recovery work for missing entries, and sampling proportion $\mathcal{R}$ is the proportion of observed entries. So our inference flow is as follows: input the partial data observed in the target year, as well as the partial missing but relatively complete historical data, and output the results of completing the missing entries for the target year by exploiting intra-disease and inter-disease correlations, and minimizing the completing error.

3 PREVALENCE INFERENCE METHODOLOGY

3.1 Methodology Overview

Our approach is based on the Missing Data Imputation method of Yoon et al. [32], which they call Generative Adversarial Imputation Nets (GAIN) to be contrasted to the traditional Generative Adversarial Network (GAN) [5].

But traditional imputation using GAIN fails to exploit inter-disease data correlations, it can only exploit intra-disease data correlations. We propose a novel approach, named CPH, to jointly fuse both intra-disease and inter-disease data correlations, to fill in the missing values in dataset. Fig. 2 depicts the overall architecture.

Firstly, all the disease matrices in $\mathcal{X}$ are taken as input, and the missing entries are initialized at the same time. Then we input them into CNN for feature extraction and get the feature matrix, then we get the corresponding mask matrix M according to the currently selected target disease (e.g., Figure 2 is based on obesity as the target disease, and inferences about different diseases can be made simply by replacing the mask matrix). Finally, the mask matrix and feature matrix are fed into the GAN model for training to obtain the complete matrix. At this point, we can further analyze the inferred results. Each of these components is described in detail as follows:

3.2 CNN-based Representation

The purpose of introducing CNN is to extract intra-disease and inter-disease data correlations. First, the missing entries of each chronic disease matrix (e.g., obesity X^1, hypertension X^2, and diabetes X^3) are initialized with different noise variables Z, Z' and Z'' obtained by sampling from either a normal or a uniform distribution. Then, we can regard the three chronic disease matrices as an image with three channels, where the two dimensions of the image represent time and space, and the number of channels indicates the different types of diseases. Finally, we can put this image into CNN-based representation, C, as shown in figure 3, to get a feature matrix X', then we define the X' by:

$$X' = C(X^1, X^2, X^3, Z, Z', Z'') \tag{2}$$

[1] https://github.com/WoodScene/Compressive-Population-Health

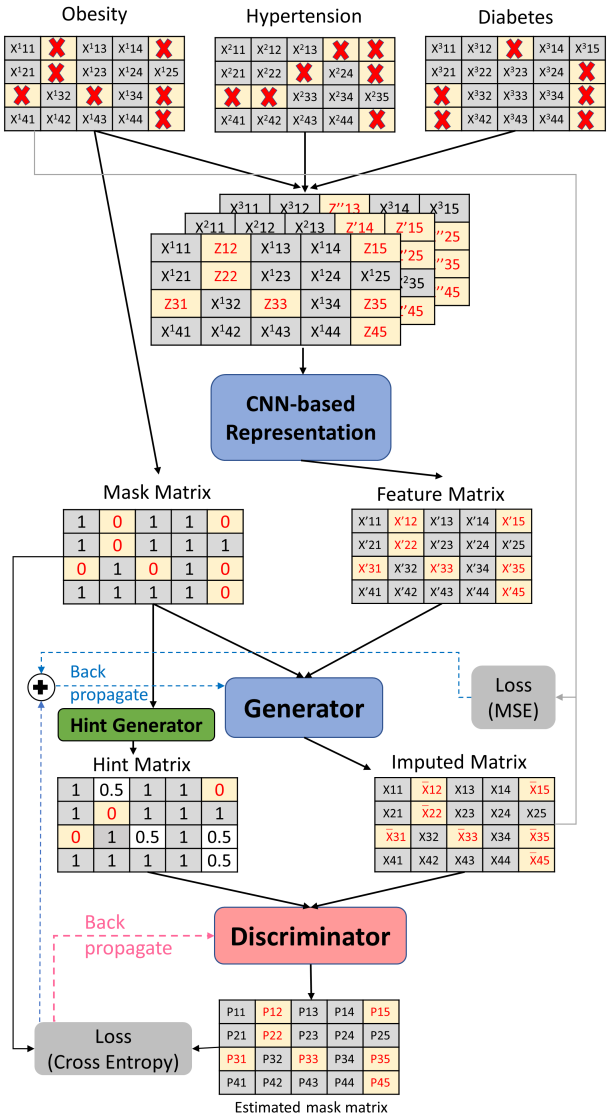


Figure 2: The CPH architecture

3.3 Generator

The generator, G, takes feature matrix X', M as input and outputs $\bar{X}$, which is a complete matrix, where the mask matrix M should change according to the target disease. For example, if obesity is chosen as the target disease, then the mask matrix should indicate which components of the obesity disease matrix are observed. Let $G : X' \times \{0,1\}^{n\times d} \to \hat{X}$ be a function. Then we define the matrices $\bar{X}, \hat{X}$ by:

$$\bar{X} = G(X', M) \tag{3}$$

$$\hat{X} = M \odot X + (1 - M) \odot \bar{X} \tag{4}$$

where $\odot$ denotes element-wise multiplication. X means the original target disease matrix. $\bar{X}$ corresponds to the matrix of imputed values (note that the output of G changes the value of each entries, even though its value is observed) and $\hat{X}$ corresponds to the completed data matrix, that is, the observations are obtained from the original target disease matrix X, and the missing values are replaced by the corresponding values in the $\bar{X}$ matrix. This setup is very similar to a standard GAN. Notice that the matrix $\hat{X}$ here is what we ultimately want to get from training.

3.4 Discriminator

The discriminator, D, like the traditional GAN framework, will act as an adversary to train G. But unlike GAN, where the output of the generator has only two choices of real or fake, under our model the output is a composite of observed and inferred entries. And D is also no longer to identify whether the input data is completely real or completely fake, but to try to distinguish which entries are observed and which are inferred, this is equivalent to predicting the mask matrix M. Finally, the discriminator is a function $D : \hat{X} \to [0,1]^{n\times d}$, with the i^{th} component of $D(\hat{X})$ corresponding to the probability that the i^{th} component of $\hat{X}$ is observed. The higher the probability value, the more likely it is that the entry is observed.

3.5 Hint

We introduce a hinting mechanism, the theoretical analysis of which can be seen in the Yoon et al. (2018) [32]. A hint mechanism is a random variable, H, taking values in a space $\mathcal{H}$, we pass H as an additional input to the discriminator and so it becomes a function $D : \hat{X} \times H \to [0,1]^{n\times d}$. By defining H differently, we can control the amount of information about M contained in H, especially if we do not provide D with "enough" information about M (e.g., we do not have a hinting mechanism at all), then several distributions that G can reproduce are optimal with respect to D. Therefore, the introduction of the hinting mechanism is necessary.

3.6 Objective

Putting it all together, we train D to maximize the probability of correctly predicting M. We train C and G to minimize the probability of D predicting M. We define the quantity $V(D, C, G)$ to be

$$\begin{aligned} V(D,C,G) = \mathbb{E}_{\hat{X},M,H}[M^T \log D(\hat{X}, H) \\ +(1-M)^T \log(1 - D(\hat{X}, H))], \end{aligned} \tag{5}$$

where log is element-wise logarithm and dependence on C and G is through $\hat{X}$.

Then, just like the standard GAN, we define the objective to be the minimax problem given by:

$$\min_{C,G} \max_{D} V(C, D, G). \tag{6}$$

We define the loss function $\mathcal{L}$: $\{0,1\}^{n\times d} \times [0,1]^{n\times d} \to \mathbb{R}$ by

$$\mathcal{L}(a,b) = \sum_{i=0}^{n} \sum_{j=0}^{d} [a_{ij} \log(b_{ij}) + (1 - a_{ij}) \log(1 - b_{ij})]. \tag{7}$$

Writing $\hat{M} = D(\hat{X}, H)$, we can then rewrite (6) as

$$\min_{C,G} \max_{D} \mathbb{E}[\mathcal{L}(M, \hat{M})]. \tag{8}$$

Both G and D are modeled as fully connected neural nets. And the models are trained using gradient descent.

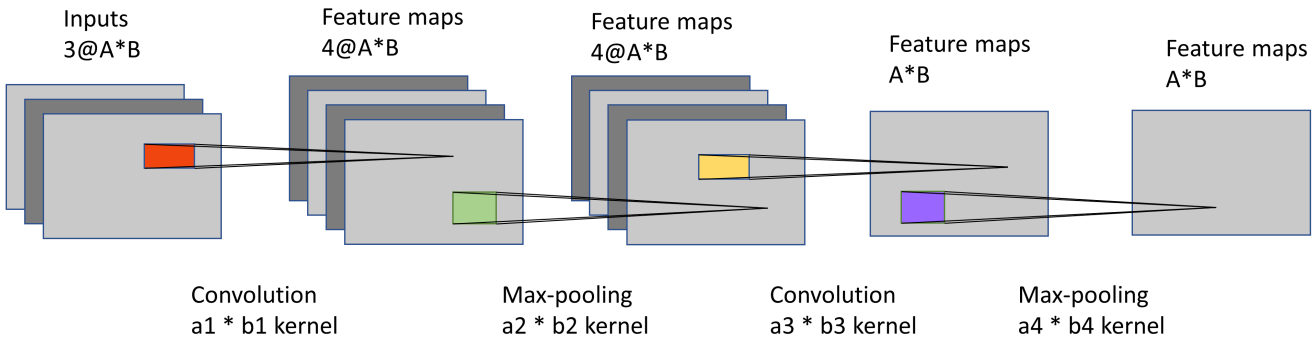


Figure 3: The architecture of CNN-based representation

4 EXPERIMENTAL EVALUATION

This section will introduce how we extensively evaluate the effectiveness of CPH based on real-world public health data. We first describe the datasets, baselines and experimental setups (Section 4.1 and Section 4.2). Then, we compare the results of different inference methods and discuss the significance of improvement (Section 4.3 and Section 4.4). Finally, we present some additional analysis on the data correlations behind the success of CPH (Section 4.5).

4.1 Datasets Description

The two datasets we use can be collected from the UK government's website without any licences:

Dataset of Ward Boundaries of London: This dataset[2] is collected from Great Britain's national mapping agency which provides the most accurate and up-to-date geographic data for government, business and individuals. In particular, the dataset includes names, shapes and codes of 630 wards in London.

Chronic Diseases Prevalence Dataset: It contains three diseases: obesity, hypertension and diabetes, which can be downloaded from The National Health Service (NHS) website[3]. The NHS publishes annual health data from 1 April to 31 March of the following year, covering 94.8% of general practices. For each type of disease, the morbidity rate is expressed as a ratio representing the number of patients on each registry to the number of all patients on the practice list. We collected the data from 2008 to 2017 of London ward level and in the raw data, roughly 30% of the regions are missing. We use dataset "2009" to represent the annual chronic disease prevalence from April 2008 to March 2009.

4.2 Baselines, Metrics, and Setups

We compare our methods to 10 baseline models, covering conventional missing data recovery methods (e.g., stKNN, CF, Linear Regression, NMF, TD), deep learning models (e.g., Auto-encoder and GAIN), and new multi-task data inference approach (DME). The detailed implementation of our method and baselines is in the Appendix part.

[2]https://data.ordnancesurvey.co.uk/

[3]https://digital.nhs.uk/data-and-information/publications/statistical/quality-and-outcomes-framework-achievement-prevalence-and-exceptions-data

Average: For each disease, take the temporal or spatial average as a complementary value.

Median: For each disease, take the temporal or spatial median as a complementary value.

stKNN: For each disease, use the average values of its k nearest spatial and temporal regions as predictions (k = 6 is the best).

CF: For each disease, User-based collaborative filtering (UCF) and Item-based collaborative filtering (ICF) are applied to generate a prediction, respectively, and the final result is the average of the two predictions.

Linear Regression: For each disease, use linear regression to predict the missing values.

NMF: For each disease, use Non-negative Matrix Factorization to predict the missing values.

TD: Construct a Tensor with three dimensions (year, grid, and disease), and use tensor decomposition to predict the missing values.

Auto-encoder: For each disease, use original auto-encoder to predict the missing values.

DME: The Deep Multimodal Encoding [13], an improved algorithm on ordinary auto-encoder. Although this work does not focus on health data, it can also simultaneously complete multiple data inference tasks while leveraging the correlations among them.

GAIN: For each disease, apply GAIN - an adaptation of generative adversarial networks (GAN) to missing data imputation.

We also conduct two variations of CPH approach to assess whether some of our detailed components are effective:

CPH$_1$−: Decrease the value of the hinting mechanism in the model to evaluate the usefulness of including a hinting mechanism.

CPH$_2$−: The simplest CNN structure is used to determine whether the introduction of CNN actually extracts intra- and inter-disease correlations.

We apply two metrics to evaluate the prediction performance: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE)

$$RMSE = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n}} \tag{9}$$

$$MAE = \frac{\sum_i |y_i - \hat{y}_i|}{n} \tag{10}$$

where $\hat{y}_i$ is an inference, y_i is the ground truth, and n is the number of inferred missing value.

We set two different experimental parameters, target year and sampling proportion. We tested the effects of completing under a variety of different target years, for example, when the target year is set to 2016, it means that we fill in the missing entries for 2016 using historical data from 2008 to 2015. And when the target year is set to 2017, it means that we fill in the missing entries for 2017 with historical data from 2008 to 2016.

The sampling proportion $\mathcal{R}$ refers to the selection of the corresponding proportion of observation entries for the target year as input to the model, where we assume that each disease is sampled in the same proportion, and the remaining observed entries will be used as a test set to validate the complementary effect of the model. For example, $\mathcal{R}$ = 0.1 means that for each disease 10% of the observed entries will be selected as the training set, and the remaining observed entries will be used as the test set (note that $\mathcal{R}$ cannot be set larger than the maximum proportion of known entries in the target year, e.g., for a disease where the proportion of all observed entries to the total is 0.7, where the maximum value of $\mathcal{R}$ should be 0.6 and at least 10% of the entries should be left as a test set). Specifically, we select a proportion of regions by a random method, at the same time ensuring that the regions selected by the different completion algorithms are the same to avoid thus affecting the experimental results.

Based on these two experimental settings, we fix one setting at a time to change the value of the other. For example, we fix the sampling proportion $\mathcal{R}$ at 0.3 and change the target year to see the results of each method. Similarly, we can also fix the target year and compare the results by changing the sampling proportion.

4.3 Inference Method Comparison

The inferred quality of the three chronic diseases under different methods is shown in Table 1, 2 and 3. From the results, we can see that CPH is superior to other baselines in completing all diseases on various settings of sampling proportions, achieving a higher inference accuracy on average from 14.8% to 9.1%. Due to space limitation, only the detailed performance on years 2016 and 2017 are presented. Then, some explanations are given to analyze the reasons our CPH model outperforms others.

Conventional missing data recovery algorithms (e.g., Linear Regression and NMF) and deep-learning-based approaches (e.g., Auto-encoder and GAN) are used to make inferences using intra-disease data correlations without considering inter-disease data correlations. One of the key advantages of CPH is that they can extract both of these two correlations automatically (through a CNN-based model) and then fuse them in the inference task for multiple diseases (through a GAN-based model), which have been experimentally verified to bring about the improvement.

For the DME baseline, although intra-disease and inter-disease correlations are also used simultaneously, they still perform worse than our CNN- and GAN-based CPH models. There are two main reasons for this, on the one hand because DME is still auto-encoder in nature, and auto-encoder is an unsupervised learning algorithm, so different data distributions may be learned during the model training, while in the CPH model we force the generator to be able to produce real data distributions by including a hinting mechanism. On the other hand, for extracting intra-disease and inter-disease correlations, the auto-encoder uses a fully connected neural network, while the CPH method uses a CNN, which is significantly superior to a fully connected neural network in extracting data features. Furthermore, CPH performed better compared to CPH_{1-} and CPH_{2-}, indicating that our further customization and improvement to CNN and GAN are in fact effective.

4.4 Significance of Improvement

In this section we show that how the improvements achieved by our CPH methods are significant for the public health with missing data problem. From the experimental results we can see that the value of RMSE and MAE are very small because morbidity rate itself is very small. From the dataset, we can see that the morbidity rates of all diseases mostly range from 0.001 to 0.1, so we normalized the value of each disease to [0,1] at first. Although the values of RMSE and MAE are small, the results of the experiment are quite significant. For example, for obesity disease, with a target year of 2017 and a sampling proportion of 0.5, the worst baseline method (CF) has a margin of error for completing 61.1%, the best baseline method (GAIN) has a margin of error of 19.0%, and the CPH method has a margin of error of only 10.5%, compared to the best baseline method (GAIN), the error rate is reduced by 8.5%.

Similarly for hypertension disease, with a target year of 2017 and a sampling proportion of 0.5, the worst baseline method (CF) has a margin of error for completing 41.2%, the best baseline method (GAIN) has a margin of error of 7.8%, and the CPH method has a margin of error of only 2.7%, compared to the best baseline method (GAIN), the error rate is reduced by 5.1%.

For diabetes disease, with a target year of 2016 and a sampling proportion of 0.5, the worst baseline method (NMF) has a margin of error for completing 51.4%, the best baseline method (DME) has a margin of error of 25.1%, and the CPH method has a margin of error of only 8.2%, compared to the best baseline method (GAIN), the error rate is reduced by 16.9%.

In addition, in the testing dataset, the real morbidity rate of obesity of the ward E05000335 in 2017 is 0.0569, the inferred value by using GAIN method is 0.0856, the MAE of this entry is 0.0287, but in the method CPH, the inferred value is 0.0604, the MAE of this entry is 0.0032. In this case, although MAE is only improved by 0.0255, it is actually improved by 45% relatively, which we believe is a significant improvement.

Now we go back to our original goal of investigating if such performance is satisfactory for real-world population health monitoring tasks. For example, we assume a maximum imputation error range of the results to be less than 15%, and then compare the minimum sampling proportion that different algorithms need to achieve this goal.

We experimentally found that CPH can sample even just 11% of the entire region to give less than 15% completing error for the remaining missing regions, but the best baseline algorithm (GAIN) needs to sample 57% of the region to satisfy the requirement. Figure 4 shows the distribution of sampling regions required to satisfy this error rate for both methods. The figure shows that the CPH method can use fewer sampling areas to perform the task of inferring population health data within a certain error range, which is important in terms of cost savings and time spent. In

Table 1: Inference quality of obesity

Methods	2016						2017					
	$\mathcal{R}$ = 0.1		$\mathcal{R}$ = 0.3		$\mathcal{R}$ = 0.5		$\mathcal{R}$ = 0.1		$\mathcal{R}$ = 0.3		$\mathcal{R}$ = 0.5	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
CF	0.1661	0.1345	0.1657	0.1340	0.1640	0.1298	0.1983	0.1625	0.2010	0.1659	0.2028	0.1702
Average(spatial)	0.1478	0.1202	0.1400	0.1153	0.1401	0.1149	0.1581	0.1310	0.1525	0.1288	0.1444	0.1229
Median(spatial)	0.1518	0.1252	0.1367	0.1110	0.1355	0.1100	0.1509	0.1233	0.1435	0.1201	0.1370	0.1150
NMF	0.1518	0.1180	0.1346	0.1064	0.1412	0.1113	0.1661	0.1331	0.1513	0.1208	0.1330	0.1054
TD	0.1403	0.1045	0.1275	0.1014	0.1250	0.1002	0.1304	0.0997	0.1221	0.0970	0.1181	0.0923
Linear Regression	0.1026	0.0763	0.0947	0.0730	0.0927	0.0687	0.1132	0.0934	0.0887	0.0714	0.0853	0.0671
Auto-encoder	0.0857	0.0616	0.0817	0.0597	0.0821	0.0597	0.0772	0.0575	0.0681	0.0520	0.0654	0.0496
stKNN	0.0794	0.0557	0.0752	0.0546	0.0732	0.0528	0.0739	0.0520	0.0632	0.0472	0.0609	0.0459
Median(temporal)	0.0830	0.0564	0.0769	0.0537	0.0760	0.0525	0.0776	0.0534	0.0662	0.0475	0.0610	0.0434
Average(temporal)	0.0788	0.0547	0.0737	0.0523	0.0728	0.0512	0.0725	0.0514	0.0615	0.0455	0.0579	0.0425
DME	0.0691	0.0525	0.0619	0.0444	0.0643	0.0435	0.0694	0.0634	0.0624	0.0459	0.0614	0.0415
GAIN	0.0948	0.0597	0.0616	0.0509	0.0580	0.0464	0.0617	0.0491	0.0507	0.0415	0.0482	0.0390
CPH_{1-}	0.0882	0.0726	0.0513	0.0393	0.0417	0.0322	0.0624	0.0498	0.0511	0.0397	0.0365	0.0288
CPH_{2-}	0.0856	0.0678	0.0718	0.0594	0.0517	0.0371	0.0608	0.0448	0.0408	0.0324	0.0369	0.0285
CPH	**0.0573**	**0.0427**	**0.0455**	**0.0352**	**0.0392**	**0.0295**	**0.0526**	**0.0411**	**0.0400**	**0.0316**	**0.0360**	**0.0281**

Table 2: Inference quality of hpertension

Methods	2016						2017					
	$\mathcal{R}$ = 0.1		$\mathcal{R}$ = 0.3		$\mathcal{R}$ = 0.5		$\mathcal{R}$ = 0.1		$\mathcal{R}$ = 0.3		$\mathcal{R}$ = 0.5	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
CF	0.1227	0.0928	0.1279	0.0973	0.1359	0.1036	0.1455	0.1094	0.1355	0.1003	0.1349	0.1027
Average(spatial)	0.1132	0.0904	0.1173	0.0937	0.1213	0.0969	0.1163	0.0935	0.1113	0.0897	0.1136	0.0903
Median(spatial)	0.1130	0.0902	0.1195	0.0954	0.1237	0.0989	0.1249	0.0998	0.1139	0.0918	0.1140	0.0906
NMF	0.0919	0.0719	0.0887	0.0713	0.0896	0.0695	0.0987	0.0784	0.0795	0.0626	0.0776	0.0615
TD	0.0782	0.0593	0.0804	0.0649	0.0798	0.0642	0.0743	0.0592	0.0780	0.0635	0.0731	0.0590
Linear Regression	0.0993	0.0925	0.0948	0.0885	0.0943	0.0880	0.0702	0.0619	0.0560	0.0524	0.0517	0.0484
Auto-encoder	0.0530	0.0392	0.0524	0.0383	0.0499	0.0369	0.0498	0.0379	0.0468	0.0365	0.0462	0.0358
stKNN	0.0461	0.0344	0.0469	0.0345	0.0436	0.0326	0.0363	0.0287	0.0324	0.0255	0.0325	0.0250
Median(temporal)	0.0472	0.0332	0.0450	0.0315	0.0438	0.0300	0.0358	0.0278	0.0267	0.0212	0.0183	0.0153
Average(temporal)	0.0448	0.0327	0.0427	0.0310	0.0403	0.0290	0.0344	0.0277	0.0257	0.0211	0.0186	0.0156
DME	0.0397	0.0298	0.0352	0.0246	0.0344	0.0214	0.0402	0.0310	0.0360	0.0261	0.0341	0.0222
GAIN	0.0371	0.0283	0.0241	0.0167	0.0258	0.0166	0.0288	0.0204	0.0201	0.0142	0.0208	0.0156
CPH_{1-}	0.0487	0.0335	0.0222	0.0170	0.0210	0.0168	0.0437	0.0362	0.0235	0.0176	0.0212	0.0157
CPH_{2-}	0.0368	0.0295	0.0212	0.0171	0.0182	0.0139	0.0249	0.0185	0.0165	0.0129	0.0169	0.0138
CPH	**0.0337**	**0.0263**	**0.0202**	**0.0158**	**0.0165**	**0.0132**	**0.0230**	**0.0167**	**0.0150**	**0.0115**	**0.0166**	**0.0129**

summary, results indicate that on the one hand, CPH, which jointly fuse both intra-disease and inter-disease data correlations, achieves significant improvements over other baseline algorithms, and on the other hand, that the idea of CPH does have an impact on advancing population health surveillance practice, saving time and cost in the data collection process.

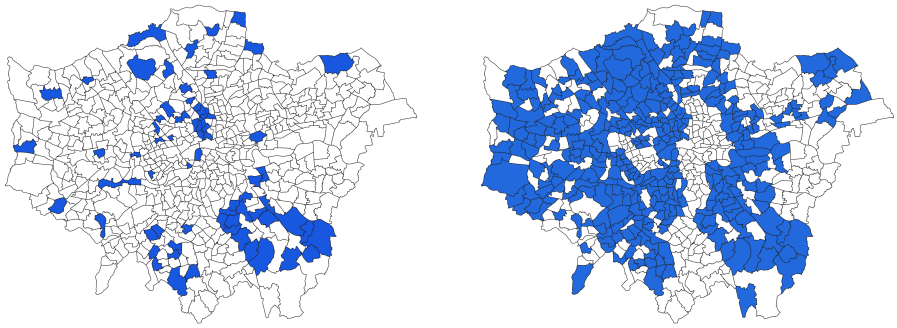
4.5 Existence of Data Correlations

This section demonstrates the existence of both intra-disease and inter-disease data correlations in population health datasets. Figure 5 shows the distribution of the prevalence of the three chronic diseases in different wards of London in 2011. From the figure we do observe some correlations among multiple disease morbidity. We can see that the incidence of diseases in neighboring wards generally reveals some similarity. This actually follows the First Law of Geography [27], "everything is related to everything else, but near things are more related than distant things."

To quantify the spatial similarity, we first calculate the Euclidean distances of all ward pairs. We adopt four difference indicators, including Arithmetic Difference (AD), Euclidean Distance (ED), Pearson Distance (PD) [26] and Cumulative Distance of Dynamic Time Warping (CDDTW) [9], to quantitatively measure the spatial correlation. The smaller the value of these four indicators, the stronger the correlation of the selected ward pairs. Therefore we look at the

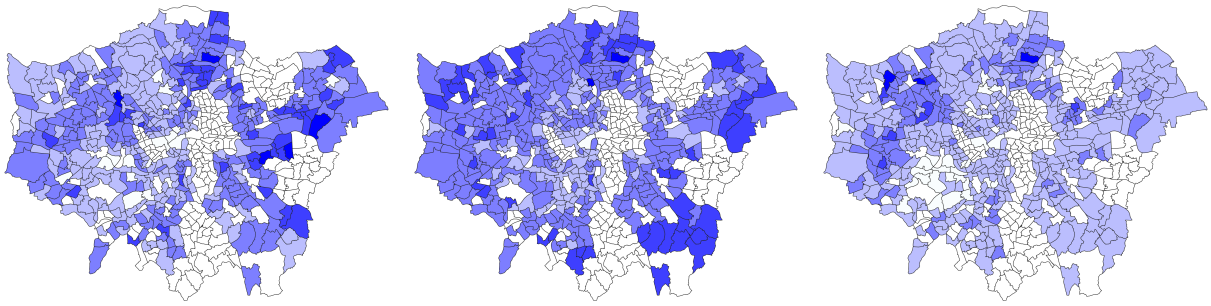
Table 3: Inference quality of diabetes

Methods	2016						2017					
	$\mathcal{R}$ = 0.1		$\mathcal{R}$ = 0.3		$\mathcal{R}$ = 0.5		$\mathcal{R}$ = 0.1		$\mathcal{R}$ = 0.3		$\mathcal{R}$ = 0.5	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
CF	0.1555	0.1171	0.1554	0.1185	0.1516	0.1151	0.2132	0.1840	0.1857	0.1572	0.1834	0.1530
Average(spatial)	0.1418	0.1120	0.1436	0.1134	0.1504	0.1199	0.2936	0.2695	0.2439	0.2250	0.2245	0.2101
Median(spatial)	0.1407	0.1106	0.1398	0.1101	0.1468	0.1174	0.2737	0.2481	0.2345	0.2149	0.2080	0.1924
NMF	0.1715	0.1363	0.1519	0.1199	0.1594	0.1268	0.2683	0.2234	0.2109	0.1758	0.1815	0.1511
TD	0.1687	0.1225	0.1468	0.1073	0.1492	0.1103	0.1513	0.1114	0.1429	0.1054	0.1213	0.0896
Linear Regression	0.0861	0.0785	0.0830	0.0748	0.0831	0.0734	0.0536	0.0416	0.0493	0.0389	0.0532	0.0483
Auto-encoder	0.1228	0.1061	0.1189	0.1022	0.1172	0.0995	0.1139	0.0979	0.1101	0.0940	0.1024	0.0856
stKNN	0.1170	0.0997	0.1096	0.0903	0.1035	0.0835	0.0878	0.0764	0.0597	0.0508	0.0411	0.0341
Median(temporal)	0.1446	0.1268	0.1405	0.1225	0.1392	0.1200	0.1253	0.1124	0.0979	0.0889	0.0735	0.0672
Average(temporal)	0.1219	0.1069	0.1183	0.1033	0.1176	0.1016	0.0996	0.0894	0.0779	0.0707	0.0585	0.0538
DME	0.1003	0.0862	0.0790	0.0678	0.0672	0.0543	0.0971	0.0834	0.0777	0.0660	0.0663	0.0534
GAIN	0.1446	0.1201	0.1054	0.0852	0.0736	0.0560	0.1658	0.1449	0.1043	0.0838	0.0325	0.0264
CPH_{1-}	0.0651	0.0568	0.0415	0.0359	0.0299	0.0239	0.0527	0.0452	0.0425	0.0325	0.0198	0.0141
CPH_{2-}	0.0806	0.0740	0.0374	0.0289	0.0379	0.0245	0.0503	0.0399	0.0327	0.0209	0.0194	0.0139
CPH	**0.0612**	**0.0237**	**0.0269**	**0.0183**	**0.0254**	**0.0181**	**0.0310**	**0.0223**	**0.0294**	**0.0202**	**0.0181**	**0.0125**



(a) CPH (b) GAIN

Figure 4: Sampling area distribution for complement error less than 15% (dark color indicates the sampled area, blank part indicates the unsampled area)



(a) Obesity (b) Hypertension (c) Diabetes

Figure 5: Morbidity rate distribution in 2011 (darker colors indicate higher morbidity rate, blank parts represent missing entries in the original data)

changes in these four metrics by increasing the distance between ward pairs. We present the empirical results of hypertension in figure 6. Then we can further make the following observations: the spatial correlations generally exist within a certain geographical scale, they are non-linear and even disappear out of a certain geographical scale.

From these observations, we can conclude that spatial correlation of morbidity must exist when the distance between two wards is not exceeding a certain threshold.

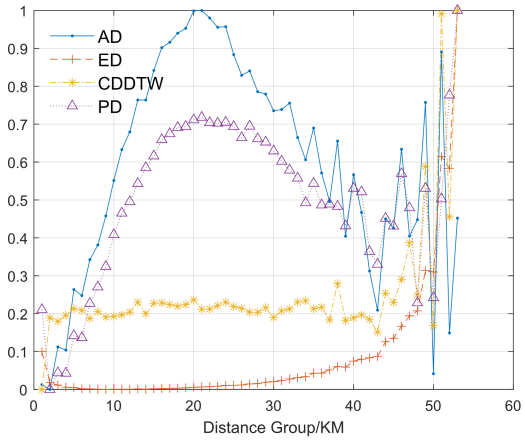


Figure 6: Morbidity difference changes as distance increases for hypertension

In addition, if we look at the three diseases in figure 5 at the same time, we will find that their prevalence distributions are quite similar. Also to further prove the inter-disease data correlations, we employ Pearson correlation analysis whose results are shown in Figure 7.

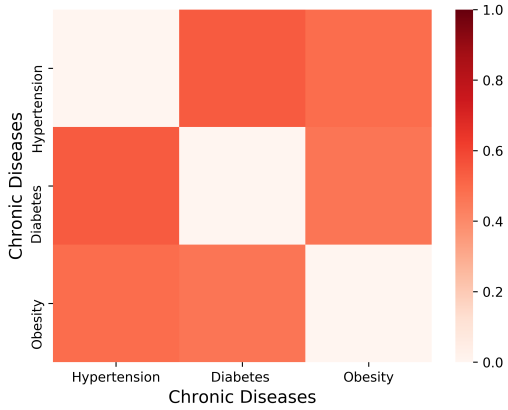


Figure 7: Correlation between different chronic diseases

Figure 7 shows that there is a strong correlation between these three chronic diseases. For example, the value of the Pearson coefficient between hypertension and diabetes is 0.54, indicating a correlation between the both. And it is supported by relevant medical literature, for example, the authors in [4] find that more than two-thirds of people with type 2 diabetes have hypertension. The studies in [11] shows that obesity is one of the main factors that leads to developing diabetes. So we can indeed exploit this intra-disease and inter-disease data correlations to improve the performance of our model.

5 RELATED WORK

In this section, we review the related research in both public health and computer science, and then compare them to our proposed approach to better clarify its novelty.

Spatial Epidemiology. Studies in the area of public health attempt to explore the factors and their effects on geographically differentiated health outcomes, including environmental variables (e.g., the air quality of surroundings), socioeconomic and demographic statistics (e.g., income), or even lifestyle choices (e.g., nutrition, diet, mobility, and sedentary). The work in [1] confirms that significant temporal and spatial correlations do exist between different chronic diseases. The authors in [7] demonstrate that the neighborhood built environment has an impact on population health. The work in [18] assessed the relationship between fast food and obesity using Instagram and Foursquare data. Mason et al [17] found that there is a strong association between high density of physical exercise facilities and reduced obesity in middle-aged adults. These studies demonstrate the existence of data correlations which we aim to leverage, and shows the feasibility of implementing CPH. However, none of these studies how to utilize these correlations to enable disease prevalence inference.

Population Health Data Inference. In recent years, an increasing number of algorithms have been used to infer missing data [15]. Recently deep learning-based models have demonstrated state-of-the-art performance in mining the massive electronic medical records data [14][16]. Multiple state-of-the-art data recovery algorithms are developed in [1] to verify that spatiotemporal correlations can be leveraged to do reliable data inference. Some studies used the structure of social networks and patterns of human mobility to study outbreak patterns of infectious diseases [19]. Others use users' posts or tweets on social networks to predict large-scale popularity patterns [6][21][28]. The authors in [30] used the population mobility patterns of metropolitan area residents to predict the prevalence of several chronic diseases in urban neighborhoods by looking at local human lifestyles. In terms of population health data inference models proposed above, although the intra-disease spatial correlations have been studied and used in the above state-of-the-art research, they do not jointly consider and incorporate the power of the inter-disease correlations.

Missing Data Recovery. From a technical perspective, completing missing data entries with spatial correlations has been studied in other domains such as environmental and traffic monitoring [10][33]. For example, the authors in [12] propose a multi-view learning method which can consider the local and global variation in temporal and spatial views to capture more information from the existing data to estimate the missing values for traffic-related time series data. The work [31] uses a similar multi-perspective based learning approach to collectively fill missing readings on Beijing air quality and meteorological data. While some others use

Compressive Sensing (CS) to environmental data recovery [23]. Compared to these studies which explore neighboring correlations for a single inference task (e.g., measurement of PM 2.5), this paper aims to explore more complex data correlations (both intra-task and inter-task correlations) to accomplish multiple inference tasks (prevalence of multiple chronic diseases). Besides, the authors in [13] propose a multimodal data fusion framework, the DME, based on deep learning techniques for missing data imputation which can exploit both intra-task and inter-task correlations. However, because the DME model is a unsupervised learning approach in nature, it cannot effectively extract correlations between multiple diseases and and combines them with intra-disease correlations, making its performance much worse than CPH.

6 LIMITATION AND DISCUSSION

In this section, we present limitations of this work followed by discussions on future work directions.

Explanations for Data Correlations. Now we have demonstrated how CPH is successful in utilizing the existence of data correlations to complete missing entries of prevalence rats for multiple chronic diseases, it would be insightful to go a step deeper into exploring the contribution of each individual factor in the dataset entries to the overall correlations. For example, we need to understand the contribution to correlation by age, gender, population migration, co-morbidity, among other factors. As future work, we aim to combine our study of CPH with state-of-the-art research work in the area of spatial epidemiology to further move this work to a deeper level.

Combination with Other Correlations. The basic intuition of CPH, which is to leverage inherent data correlations to perform inference, can be extended beyond intra-disease spatial correlations and inter-disease correlations, seeking to improve inference accuracy. Other such correlations that we aim exploit in the future include multi-source urban big data [3] (also known as smart city data) which is becoming increasingly available, e.g., population density, mobility, traffic data, education and economic status, age distribution, population participation in events and activities, air quality measures, among others. Some of these data sources may have an impact on the population health status. In future work, we will explore how to use such additional correlations to further improve the data recovery model.

Missing Prevalence Completion for All Years. In this paper, we set our goal as to constructing a full disease prevalence map of the current year based on the correlations learned from historical data in previous years, so that a timely public health monitoring map can be established before sharing them on the Web for research. Here, in our experimental evaluation, the prevalence rates data in previous years, though with some missing entries, are relatively complete, so that CPH can effectively extract the correlations and achieve the data completion accurately for the current year even with high data missing rates. However, in some circumstances (e.g, in developing countries), such a relatively complete historical prevalence are not available, and a more complicated and meaningful task is to build a model CPH+ to infer the missing entries for all years (both the current and historical years). In this case, how to effectively learn these two correlations from sparse historical data would be a more challenging research problem. For example, we can consider to transfer the extracted correlations in developed countries to developing ones based on transfer learning strategies, which is an interesting research direction to explore in the future.

Sustainability of CPH. Although the experiment shows that CPH is capable of learning both intra- and inter-disease data correlations from the historical data and then utilizing them in the prevalence rate completion task of the current year, we have not explore yet how to make CPH to be sustainable year after year. As time goes by, the historical data will contain both collected and inferred data entries, and there is a risk that the error in the inferred entries will be propagated and accumulated when adopting CPH on the new year. Therefore, it is an interesting and challenging direction in the future to develop a sustainable CPH. One possible idea is to estimate the reliability of inferred entries, which will be regarded as weights in the prevalence inference model. The weights will be dynamically updated before CPH will be adopted on a new year as more ground-truth data is available. More interestingly, the inferred data entries may be calibrated to more accurate ones.

7 CONCLUSION

This paper proposes a deep-learning-based approach, we call Compressive Population Health (CPH), to infer and recover missing entries of population health prevalence rates of multiple chronic diseases. By reliably recovering such missing data, a full picture of the public health surveillance can be built and used. The data recovery is enabled by exploiting intra-disease and inter-disease correlation opportunities. Specifically, we first proposed a Convolutional Neural Network (CNN) based approach to extract and model both of these two types of correlations, and then adopted a Generative Adversarial Network (GAN) based prevalence inference model to jointly fuse their combined effect. We extensively evaluated the inference model based on real-world public health datasets, and the results demonstrated that CPH outperforms other baselines in various settings, and with a much improved accuracy.

8 ACKNOWLEDGEMENT

This work was supported by NSFC (National Natural Science Foundation of China) under Grant No. 61872010, the National Science and Technology Major Project (No. 2018ZX10201002), and the Project 2019BD005 supported by PKU-Baidu fund.

REFERENCES

[1] Dawei Chen, Jiangtao Wang, Wenjie Ruan, Qiang Ni, and Sumi Helal. 2020. Enabling Cost-Effective Population Health Monitoring By Exploiting Spatiotemporal Correlation: An Empirical Study. (2020).
[2] Sarah Curtis. 2004. Health and inequality: geographical perspectives. *health inequality geographical perspectives* 1 (2004), 94.
[3] Peter Dwyer, Lisa Scullion, Katy Jones, Jenny McNeill, and Alasdair B.R. Stewart. 2020. Work, welfare, and wellbeing: the impacts of welfare conditionality on people with mental health impairments in the UK. *Social Policy and Administration* 54, 2 (March 2020), 311–326. https://doi.org/10.1111/spol.12560
[4] Ele Ferrannini and William C Cushman. 2012. Diabetes and hypertension: the bad companions. *The Lancet* 380, 9841 (2012), 601–610.
[5] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. *Advances in Neural Information Processing Systems* 3 (2014), 2672–2680.
[6] Sangeeta Grover and Gagangeet Singh Aujla. 2015. Twitter data based prediction model for influenza epidemic. In *International Conference on Computing for Sustainable Global Development.*

[7] Apinan Hasthanasombat and Cecilia Mascolo. 2019. Understanding the Effects of the Neighbourhood Built Environment on Public Health with Open Data. In *The World Wide Web Conference*.
[8] Karen, Barnett, Stewart, W, Mercer, Michael, Norbury, Graham, Watt, and Sally and. 2012. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *Lancet* (2012).
[9] Eamonn Keogh and Chotirat Ann Ratanamahatana. 2005. Exact indexing of dynamic time warping. *Knowledge & Information Systems* 7, 3 (2005), 358–386.
[10] Linghe Kong, Mingyuan Xia, Xiao-Yang Liu, Guangshuo Chen, Yu Gu, Min-You Wu, and Xue Liu. 2013. Data loss and reconstruction in wireless sensor networks. *IEEE Transactions on Parallel and Distributed Systems* 25, 11 (2013), 2818–2828.
[11] The Lancet. 2017. Obesity and diabetes in 2017: a new year. *Lancet* 389, 10064 (2017), 1.
[12] Linchao Li, Jian Zhang, Yonggang Wang, and Bin Ran. 2018. Missing Value Imputation for Traffic-Related Time Series Data Based on a Multi-View Learning Method. *Intelligent Transportation Systems, IEEE Transactions on* (2018).
[13] Zuozhu Liu, Wenyu Zhang, Shaowei Lin, and Tony Q. S. Quek. 2017. Heterogeneous Sensor Data Fusion By Deep Multimodal Encoding. *IEEE Journal of Selected Topics in Signal Processing* 11, 3 (2017), 479–491.
[14] Liantao Ma, Junyi Gao, Yasha Wang, Chaohe Zhang, Jiangtao Wang, Wenjie Ruan, Wen Tang, Xin Gao, and Xinyu Ma. 2020. AdaCare: Explainable Clinical Health Status Representation Learning via Scale-Adaptive Feature Extraction and Recalibration. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.
[15] Liantao Ma, Xinyu Ma, Junyi Gao, Xianfeng Jiao, Zhihao Yu, Chaohe Zhang, Wenjie Ruan, Yasha Wang, Wen Tang, and Jiangtao Wang. 2021. DistCare: Distilling Knowledge from Publicly Available OnlineEMR Data to Emerging Epidemic for Prognosis. In *The Web Conference (WWW)*.
[16] Liantao Ma, Chaohe Zhang, Yasha Wang, Wenjie Ruan, Jiangtao Wang, Wen Tang, Xinyu Ma, Xin Gao, and Junyi Gao. 2020. ConCare: Personalized Clinical Feature Embedding via Capturing the Healthcare Context. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.
[17] Kate E Mason, Neil Pearce, and Steven Cummins. 2017. Associations between fast food and physical activity environments and adiposity in mid-life: cross-sectional, observational evidence from UK Biobank. *The Lancet Public Health* 3, 1 (2017).
[18] Yelena Mejova, Hamed Haddadi, Anastasios Noulas, and Ingmar Weber. 2015. FoodPorn: Obesity Patterns in Culinary Interactions. In *International Conference on Digital Health*.
[19] Lauren Ancel Meyers. 2007. Contact network epidemiology: Bond percolation applied to infectious disease prediction and control. *Bull. Amer. Math. Soc.* (2007).
[20] OECD. 2015. *Health Data Governance*. 200 pages. https://doi.org/https://doi.org/10.1787/9789264244566-en
[21] Michael J Paul and Mark Dredze. 2011. You are what you tweet: Analyzing twitter for public health. In *Fifth international AAAI conference on weblogs and social media*. Citeseer.
[22] John R Pleis and Jacqueline W Lucas. 2014. Summary Health Statistics for US Adults: National Health Interview Survey. *Vital Health Stat* 260, 260 (2014), 1–161.
[23] Giorgio Quer, Riccardo Masiero, Gianluigi Pillonetto, Michele Rossi, and Michele Zorzi. 2012. Sensing, Compression, and Recovery for WSNs: Sparse Signal Modeling and Monitoring Framework. *IEEE Transactions on Wireless Communications* 11, 10 (2012), 3447–3461.
[24] Jeremy A Rassen, Dorothee B Bartels, Sebastian Schneeweiss, Amanda R Patrick, and William Murk. 2018. Measuring prevalence and incidence of chronic conditions in claims and electronic health record databases. *Clinical Epidemiology* Volume 11 (2018), 1–15.
[25] Lincoln A. Sargeant, Nina Heaps, and Penny Miller. 2007. Dealing with incomplete and inaccurate data in public health: case study of a health equity audit of health visiting services. *Journal of Public Health* 29, 3 (05 2007), 321–321. https://doi.org/10.1093/pubmed/fdm021 arXiv:https://academic.oup.com/jpubhealth/article-pdf/29/3/321/4511049/fdm021.pdf
[26] Mary C. Seiler and Fritz A. Seiler. 2010. Numerical Recipes in C: The Art of Scientific Computing. *Risk Analysis* 9, 3 (2010), 415–416.
[27] Waldo R. Tobler. 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography* 46, 2 (1970).
[28] Shikha Verma, Younghee Park, and Mihui Kim. 2017. Predicting Flu-Rate Using Big Data Analytics Based on Social Data and Weather Conditions. *Advanced Science Letters* (2017).
[29] Marieke Verschuuren and Hans Van Oers. 2019. *Population Health Monitoring Climbing the Information Pyramid: Climbing the Information Pyramid*.
[30] Yingzi Wang, Xiao Zhou, Anastasios Noulas, Cecilia Mascolo, Xing Xie, and Enhong Chen. 2018. Predicting the Spatio-Temporal Evolution of Chronic Diseases in Population with Human Mobility Data. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, 3578–3584. https://doi.org/10.24963/ijcai.2018/497
[31] Xiuwen Yi, Yu Zheng, Junbo Zhang, and Tianrui Li. 2016. ST-MVL: Filling Missing Values in Geo-Sensory Time Series Data. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (New York, New York, USA) *(IJCAI'16)*. AAAI Press, 2704–2710.
[32] Jinsung Yoon. 2018. GAIN: Missing Data Imputation using Generative Adversarial Nets. (2018).
[33] Zhu, Yanmin, Li, Zhi, Hongzi, and Minglu. 2013. A Compressive Sensing Approach to Urban Traffic Estimation with Probe Vehicles. *IEEE Transactions on Mobile Computing* (2013).

A DETAILS OF EXPERIMENTAL SETTINGS

Experimental environment: the programming language is python 3.7, the code is based on Tensorflow 1.14.0.

A.1 CPH Model Implementation

Our model is an improvement on the GAIN model, where the GAN part of the code is in paper [32]. We added CNN to their model and also adjusted the input format of the model. The network structure of the CNN is shown in Figure 3. The two-layer convolution and pooling operation is used, where the hyperparameters are as follows: the parameter of the filter in the first convolution layer operation is 1*4, number of 3, step 1, and the maximum pooling parameter is 1*4, step 1. The parameter of the filter in the second convolution layer operation is 1*4, number of 1, step 1, and the maximum pooling parameter is 1*4, step 1. The parameters in the GAN model have been experimented with and we still use the settings recommended in the original paper.

A.2 Baseline Implementations

Our principles for setting the hyper-parameters for each baseline model are as follows. If the hyperparameter settings are available in the original paper, we will use the recommended settings. Otherwise, the hyperparameters of the baseline model will be fine-tuned by the grid search strategy.

NMF: We set the embedding dimension as 5, the initial learning rate is 0.1, and decreases as the number of training sessions increases, stopping training when the change in loss is less than 1e-3.

TD: We set the dimensions of the core tensor to 10, 10, 10, with an initial learning rate of 0.1 and decreasing as the number of training sessions increases, stopping training when the change in loss is less than 3e-4.

Auto-encoder: We set four fully connected hidden layers, each with 32 neurons, a learning rate of 5e-3, a regular term coefficient of 1e-2, and a number of training rounds of 1e5.

DME: We set up four hidden layers, each with 96 neurons, a learning rate of 5e-3, a regular term coefficient of 5e-1, and a number of training rounds of 1e5, the network structure is as shown in paper [13].

GAIN: We use the settings described in the code of the paper [32].

Other baselines (e.g., linear regression, mean, and average) do not have hyperparameter settings and their results are only related to the data itself.