Quality-Guaranteed and Cost-Effective Population Health Profiling: A Deep Active Learning Approach

Chen, L., Wang, J. & Thakuriah, P

DOI 10.1145/3617179
ISSN 2691-1957

Publisher: Association for Computing Machinery (ACM)

Quality-Guaranteed and Cost-Effective Population Health Profiling: A Deep Active Learning Approach

LONG CHEN and JIANGTAO WANG *, Center for Intelligent Healthcare, Coventry University, UK
PIYUSHIMITA (VONU) THAKURIAH, Rutgers Urban and Civic Informatics Lab, Rutgers University, New Brunswick, New Jersey, US

Reliability and cost are two primary consideration for profiling population-scale prevalence (**PPP**) of multiple None Communicable Diseases (**NCDs**). In this paper, we exploit intra-disease and inter-disease correlation in different traditionally-sensed-areas (**TS-A**) to reduce the required number of the profiling task allocated without compromising the data reliability. Specifically, we propose a novel approach called Compressive Population Health TS-A Selection (***CPH-TS***), which blends the state-of-the-art profile inference, data augmentation and active learning in a unified deep learning framework. It can actively select a minimum number of TS-A regions for profiling task allocation in each profiling cycle, while deducting of the missing data of the unprofiled regions with a probabilistic guarantee of reliability. We evaluate our approach on real-world prevalence datasets of London, which shows the effectiveness of *CPH-TS*. In general, *CPH-TS* assigned 11.1-27.3% fewer tasks than baselines, assigning tasks to only 34.7% of the sub-regions while the profiling error below 5% for 95% of the cycles.

CCS Concepts: • **Applied computing** → **Health informatics**.

Additional Key Words and Phrases: Profiling of Prevalence, Spatio-temporal Correlations, Generative Adversarial Network, Convolutional Neural Networks (CNN).

1 INTRODUCTION

Non-communicable diseases (NCDs), such as heart attack, hypertension, and cancer, is one of the most common causes of death in UK. Quantifying and understanding the NCDs patterns and trends is the central task for the healthcare authorities to manage the health intervention programs. Generally speaking, one can handle this task via population health surveillance, which is a institutional sensing that collects information of the health status of a population.

At the core of population health surveillance is the so-called profiling population-scale prevalence (**PPP**), which aims to profile the morbidity rate of multiple NCDs. There are two popular ways to conduct **PPP**, which are clinical-record integration and residential survey. Unfortunately, none of them is a trivial task for NCDs surveillance [24]. The former entails the access of private health data, which is sensitive and thus needs some extra work to ensure the anonymity when conducting the data integration. The latter requires the recruitment of a residents group, which acts as the representative for the entire population. Each resident will be assigned with an interview or questionnaires, which is time consuming and costly for the process of survey administration.

*Jiangtao Wang is the corresponding author.

Authors' addresses: Long Chen, ad8579@coventry.ac.uk; Jiangtao Wang, ad5187@coventry.ac.uk, Center for Intelligent Healthcare, Coventry University , P.O. Box 412, Coventry, UK, CV1 5RW; Piyushimita (Vonu) Thakuriah, Rutgers Urban and Civic Informatics Lab, Rutgers University, New Brunswick, New Jersey, Suite 400, Civic Square Building, Rutgers, US, p.thakuriah@rutgers.edu.

2637-8051/2023/8-ART $15.00
https://doi.org/10.1145/3617179

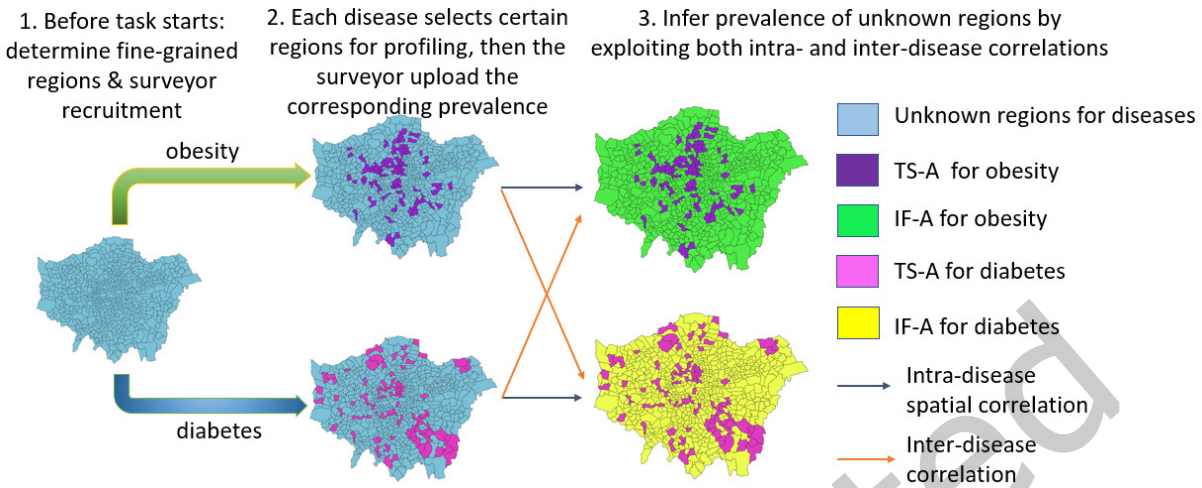

Fig. 1. Basic Vision of TS-A selection for Compressive Population Health of London in 2014[4]

[12] proposed a new compressive population health paradigm, which can infer the prevalence rate of unknown regions automatically. However, it only focuses on the data reconstruction phase without considering how to selecting the most salient TS-A. Therefore, in this work, we aim to reduce the cost to the maximum extent for the data reconstruction by employing deep active learning techniques. We propose a novel health data science paradigm called Compressive Population Health [12] Traditionally-Sensed-Areas Selection (*CPH-TS*) , which has been made publicly available [1]. In addition to achieve the major goal of cost reduction, we also make sure that the profiled information by our new approach is reliable by using Bayesian inference with a probabilistic guarantee of reliability. The expected transformative outcome is to benefit the public health authorities (e.g., NHS [2] and PHE [3]) in reducing the economic burden on the population health surveillance tasks. As Figure 1 shows, our basic vision is that, for each target disease (e.g., obesity, hypertension, and diabetes), *CPH-TS* only selects its "best" subset of regions (called Traditionally-Sensed-Areas, TS-A for short) where public health administrators still profile the prevalence rate through traditional method (either hospital-visit-based data integration or survey-based approach). Then, *CPH-TS* uses prevalence rate measured from TS-A to perform inference on the un-selected regions (called "Inferred Areas", IF-A for short). The prevalence rate inference is facilitated by exploiting the inherent data correlations extracted from historical data in multiple open-access public health datasets and evidence from epidemiology research.

To address the aforementioned challenges, we proposed a novel approach, namely Compressive Population Health Traditionally-Sensed-Areas Selection (*CPH-TS*).

In order to achieve the **PPP** task above-mentioned, we need to address research questions as follows:

Challenge A: How many and which regions should be chosen as TS-A for each disease?

In each year, we need to minimize the number of the allocated TS-A while ensuring the data reliability. In order to find this minimum TS-A collection, we need to identify the salient regions whose prevalence is informative to deduce the prevalence of other regions to the maximum extent. However, how to identify the salient regions in

[1]https://github.com/long4coventry/CPH-TS/
[2]https://www.nhs.uk/
[3]https://www.gov.uk/government/organisations/public-health-england
[4]This figure has been referred from our previous conference paper [12]

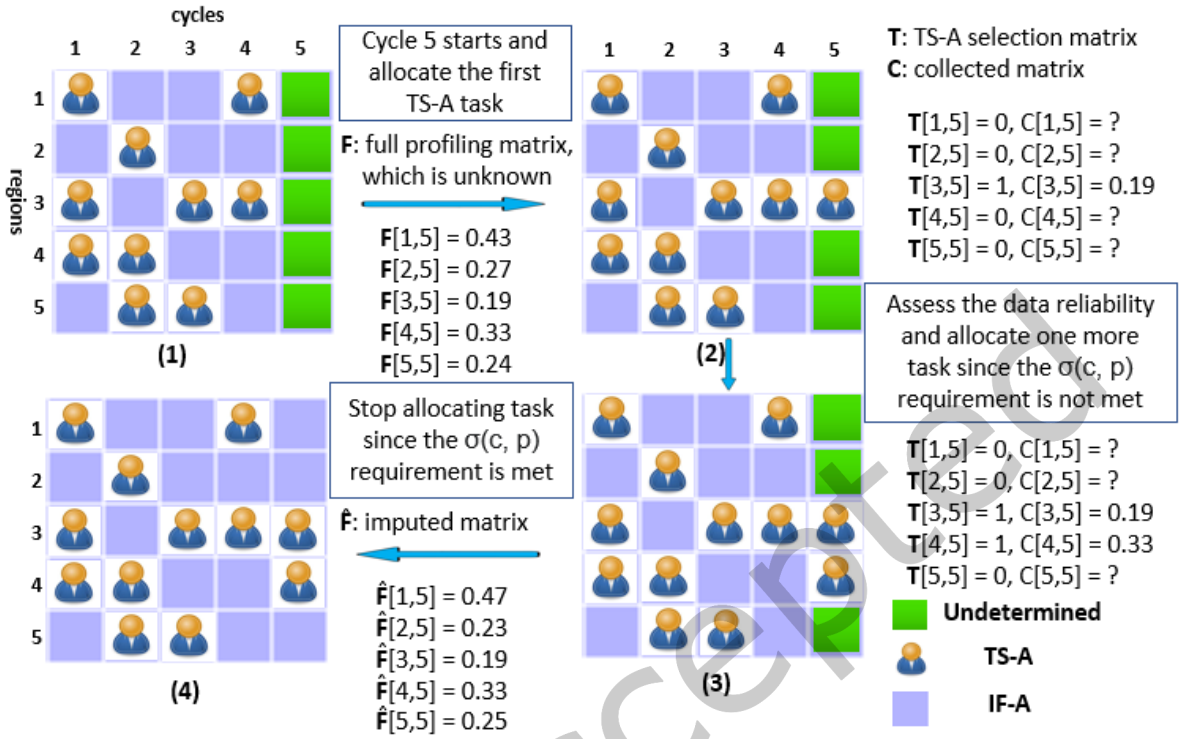

Fig. 2. The workflow of *CPH-TS*

an incremental manner is not trivial, since without foreseeing the true prevalence of a region, it is difficult to tell how much that value can enhance the data reliability.

Challenge B: How to estimate the data reliability online throughout a prevalence profiling task without knowing the true prevalence of the unprofiled regions?

Since the true prevalence of the unprofiled regions is unknown, we cannot calculate the data reliability directly by comparing the deduced values with the actual ones. Hence, it is important to estimate such profiling data reliability online in each profiling cycle. Furthermore, as the estimated data reliability intrinsically has certain deviation from the actual prevalence rate, it is impractical to require all the profiling cycles to meet a predefined error bound. Therefore, we need to define the data reliability requirement for the entire **PPP** task instead of merely having an error bound for each profiling cycle.

Given the challenges and caveats described above, the contributions of the paper are listed as follows:

1) We proposed a cost-effective population health profiling approach through multi-task active learning, which can jointly consider both the uncertainty and diversity when selecting the optimal regions to conduct profiling tasks for multiple diseases.

2) To control the profiling cost and ensure quality, we present a novel profiling data quality metric to evaluate the reliability of an T-SA task in each cycle, as well as a novel Bootstrapping Bayesian-Inference(BBI) method to learn the stopping criterion.

3) The existing Generative Adversarial Imputation Nets (GAIN) only exploits intra-disease correlation (spatial correlation), therefore we proposed a multi-task learning framework for data imputation, which extracts and

exploits both intra-disease correlation (spatial correlation) and inter-disease correlation (multimorbidity) for data imputation. Notice that the direct adoption of GAIN can only explore intra-disease correlaiton, whereas *CPH-TS* can exploit both intra-disease and inter-disease correlations.

4) We undertake rigorous experimental assessments on real-world NCDs datasets from London in order to demonstrate *CPH-TS*'s efficacy. In general, *CPH-TS* assigned 11.1%-27.3% less tasks than baselines, with profiling tasks assigned to only 34.7% of sub-regions and a profiling error of less than 5% for 95% of cycles.

2 RELATED WORK

2.1 TS-A Selection in **PPP**

Most of the TS-A selection in **PPP** was designed around the metric of spatial coverage. For example, Clements et al., [9] used a stratified cluster random sampling to optimize the survey task distribution of schistosome infection. Similarly, Gitonga et al., [14] designed a non-probability sampling to guarantee a sensible spatial spread of survey tasks regarding Plasmodium infection. Linli more recently proposed L2MM [22], a sampling technique based on deep learning for handling low-quality GPS data. Alternatively, survey tasks are often selected to maximize the spatial coverage over a given area regarding its shape and existing statistics. For example, Groenigen [37] proposed an iterative process to tackle the task allocation of soil sampling, which attempt to minimise the distance among all the data points so as to obtain a training dataset with an uniform distribution of sites with any given shape. Unlike the existing work, where coverage ratio is regarded as the reliability metric, which cannot be directly applied to active learning tasks since it only takes account of the diversity in its acquisition function. In this work, we use the overall profiling error to evaluate the data reliability, based on which we want to minimize the number of the selected tasks so that the **PPP** organizers can save budget. Notice that [40] also proposed an overall profiling error scheme to facilitate active learning in the domain of compressive crowd sensing, which is somewhat similar to our idea. However, in addition to the domain difference, their acquisition function is purely based on a traditional active learning method, e.g., query-by-committee [4], whereas ours is based on BatchBALD, which is a state-of-the-art deep learning algorithm.

2.2 Population Health Data Inference

There have been a large number of studies strive to infer the missing data. For example, [2, 33, 38] proposed various data imputation methods to recover the missing data for NCDs surveillance. However, they are largely relied on the traditional matrix recovery models. More recently, the advances in deep learning models have shown state-of-the-art performance in inferring the enormous amount of electronic records [28, 30]. Ma et al.,[29] proposed several data recovery methods that exploits the spatio-temporal correlation to achieve reliable data inference. Some studies take advantage of the geographical patterns of social network [16] or sensor data [25] to infer the patterns of disease outbreak. In [41], the authors investigate the mobility patterns of city residents to infer the prevalence rate of multiple chronic NCDs, which is somewhat similar to our idea. However, they only focused on spatial correlation of the disease, while our model also harness the power of inter-disease correlation. CPH [12] is a recent research that takes into account both inter- and intra-disease correlations, which is comparable to our approach. However, their work is entirely focused on high-quality data reconstruction, whereas ours is primarily concerned with the trade-off between reliability and cost through the application of deep active learning.

2.3 Deep Active Learning

The batch-based sampling strategy is the foundation of Deep Active Learning(DeepAL) [31], since the traditional one-by-one sampling strategy can only make a marginal change on the training space in each iteration, which is ineffective when it comes to large dataset. In addition, batch sampling can largely reduce the profiling time as

the survey tasks were done in a parallel, which in turn save the general cost. The naive approach of DeepAL is to select a batch of samples by using the original one-by-one sampling strategy. For example, [13, 21, 34] combined batch acquisition with Bayesian active learning for disagreement (BALD) [20], which simply selects top k samples with the highest disagreement score. While effective, DeepAL along this line often leads to a training dataset comprise informative but very similar samples. The information extracted from these similar samples is essentially the same, which wasted the valuable surveying resources. In addition, this sampling method assumes that samples are independent to each other, and thus ignores the correlation among samples. To address the above limitation, BatchBALD [23] is proposed, which incorporates the correlation among samples by estimating the mutual information of the batch between the sample embeddings and model parameters.

In addition, traditional active learning methods consider query strategy (i.e., TS-A selection in this paper) and training of the predicative model as two separate problem. While effective, models along this line often lead to sub-optimal result. This is because query strategies are based on fixed feature representation, but in most deep learning scenarios, feature representation is dynamically updated in the training process of the predicative models, which leads to two following drawbacks. Firstly, as these two problems correspond to a respective loss function, the dependency information between these two problems is completely ignored which often leads to divergence issue. Secondly, waiting for the output from the predicative model would incur some extra computional time. To address the problem above, researchers have proposed the end-to-end DeepAL models that use a single unified loss function for the entire active learning process. For example, LLAL [43] uses a comprehensive loss based on both target loss of active learning and the loss-prediction loss of deep learning model. MCDAL [8] combines both stages with an integrated classifier. However, end-to-end deep active learning models start with a small amount of labeled samples while the training process of deep learning models often relies on a large amount of labeled data. Therefore, we argue that a more principled DeepAL design also needs to exploit a large amount of samples to resolve this problem. The cost effective active learning (CEAL) [39] expands the original training dataset with unlabelled samples that have a high prediction score. However, the studies along this line assumes the availability of a large amount of unlabeled samples, which is not always feasible in the real world. Hence, Generative Adversarial Neural Network (GAN) [15] is proposed for data augmentation. A typical example is GAAL [46], which uses the synthesized samples with more information to enrich the original training samples. This is arguably the seminal work that uses GAN for active learning. But it is still based on a two-step approach. To the best of our knowledge, *CPH-TS* is allegedly the first end-to-end DeepAL model that exploits the GAN-based deep learning framework.

3 PROBLEM STATEMENT

To formally define the **PPP** task, we define the concepts about the profiling and selection matrices (c.f., Figure 2).

Definition 1. Full Disease Matrix. For a region-centric **PPP** task of m regions and c profiling cycles, its full disease matrix is given as $F_{r\times n}$, where each entry $F[i, j]$ is the true profiling data of region i in cycle j.

Definition 2. Region-Selection Matrix. In a region-selection matrix $T_{r\times n}$, each entry $T[i, j]$ denotes whether or not the corresponding entry in the full disease matrix $F[i, j]$ is selected for profiling: if region i is selected for profiling in cycle j, then $T[i, j] = 1$, otherwise $T[i, j] = 0$.

Definition 3. Collected Disease Matrix. A collected disease matrix $C_{r\times n}$ records the actual collected profiling data:

$$C = F \otimes T \tag{1}$$

where the $\otimes$ operator conducts the *Hadamard product* [19]] of two matrices.

Definition 4. Disease Matrix Imputation Algorithm. A disease matrix imputation algorithm **R** will impute a full disease matrix $F_{r\times n}$ from the collected disease matrix $C_{r\times n}$:

$$R(C_{r\times n}) = \hat{F}_{r\times n} \approx F_{r\times n} \tag{2}$$

where r is the number of region and n is the number of cycles. Now, we define the overall profiling error that can directly evaluate the data reliability.

Definition 5. Overall profiling Error. It measures the difference between the imputed full disease matrix $\hat{F}$ and the true disease matrix F. We calculate the overall profiling error of each profiling cycle separately. For profiling cycle k, the overall profiling error is defined as:

$$\mathcal{E}_k = \text{error}(\hat{F}[:,k], F[:,k]) \tag{3}$$

where $F[:,k]$ denotes the true profiling values of all the r TS-A in cycle k, and $\hat{F}[:,k]$ represents the imputed values by using the matrix completion algorithm R.

Definition 6. Reliability metric $\sigma(\epsilon, p)$. For a **PPP** task of n profiling cycles, it satisfies the reliability requirement of $\sigma(\epsilon, p)$, iff

$$|\{k \mid \mathcal{E}_k \le \epsilon, 1 \le k \le n\}| \ge n \cdot p \tag{4}$$

where p is a the probability that determines the least percentage of cycles whose error should be smaller than ϵ. Naturally, we want a **PPP** task to make the overall profiling error lower than ϵ in all ($p = 1$) the cycles. However, this scenario is impossible in the real world since knowing the accurate profiling error $\mathcal{E}_k$ beforehand is infeasible. Hence, we can only select a reasonable p value in terms of statistics (for example, 90% or 95%) to ensure that the overall profiling error falls within the error bound ϵ for most of the cycles.

Now, the **PPP** task can be formally summarized as the following problem. Given a **PPP** task with r regions and n cycles, and a disease matrix imputation algorithm R, we aim to obtain the minimum subset of profiling TS-A during the whole **PPP** task process (i.e., minimize the non-zero cells in the TS-A matrix T), while keeping the overall profiling errors of $n \cdot p$ cycles below the predefined error bound $\sigma(p, \epsilon)$. The matrix completion algorithm R aims to impute a full disease matrix $\hat{F}_{r\times n}$ from the collected disease matrix $C_{r\times n}$.

$$\begin{aligned} &\min \sum_{i=1}^{r}\sum_{j=1}^{c} T[i,j] \\ &\text{s.t., } |\{k \mid \mathcal{E}_k \le \epsilon, 1 \le k \le n\}| \ge c \cdot p \\ &\text{where } \mathcal{E}_k = \text{error}(\hat{F}[:,k], F[:,k]) \\ &\hat{F} = \mathcal{R}(C), C = F \otimes T \end{aligned} \tag{5}$$

When F is given, we can identify the optimal T by enumerating all the cells (regardless of the computational cost). However, in real world, F is mostly unknown, which makes it a thorny problem for the following two reasons: (1) $\mathcal{E}_k$ cannot be directly calculated, and (2) the TS-A selection is a monotonic process (i.e., we cannot get $F[i,j]$ until we set $T[i,j] = 1$, where this operation cannot be retracted to save the costs). To take account of these factors, we propose *CPH-TS*, which uses an iterative pipeline to select TS-A for profiling in each cycle, which will be explained in details in the next section.

4 THE OVERVIEW OF *CPH-TS*

In this section, we introduce the architecture of *CPH-TS*. To begin with, *CPH-TS* assumes that there are abundant participants in each region all the time, which makes it possible to collect prevalence rate from any target TS-A regions. In real-world scenarios, this may not be the case and we will discuss how to relax this assumption in the future work.

Fig. 2 shows the pipeline of *CPH-TS*. In each cycle, *CPH-TS* selects the next salient TS-A for profiling and waits for the health experts to collect the prevalence data(e.g., clinic records or health surveys) in that TS-A, until the estimated data reliability satisfies the predefined $\sigma(\epsilon, p)$ requirement. Then, the task allocation formally ends and the missing data are filled with IF-A in one profiling cycle. Suppose the target TS-A contains five regions and the fifth profiling cycle is in the current stage; to start with, no profiling data is collected in cycle five. The pipeline of *CPH-TS* is summarized as follows:

(1) *CPH-TS* pinpoints the most salient region (Figure 2-2, region 3, $T[3,5] = 1$) and select a TS-A task for region 3 to complete. The health expert conduct the **PPP** task and returns the data to the *CPH-TS* (Figure 3-2, $C[3,5] = F[3,5] = 0.19$).
(2) After *CPH-TS* acquires the profiling data of region 3, it evaluates whether the data reliability meets the predefined $\sigma(c, p)$ requirement. Let the result be no, then *CPH-TS* continues searching the next salient region (region 5, $T[4,5] = 1$) to allocate another TS-A task (Figure 2-3, $C[4,5] = F[4,5] = 0.33$).
(3) After completing the TS-A tasks from region 3 and 4 in cycle 5, *CPH-TS* evaluates whether the data reliability meets the predefined $\sigma(c, p)$ again. If the updated result is yes, CCS-TA stops further task allocations for cycle 5 and imputes all the missing data of the undetermined regions (Figure 2-4, $\hat{F}[1,5]$, $\hat{F}[2,5]$, $and\hat{F}[5,5]$are imputed).

5 DETAILED DESIGN OF *CPH-TS*

In this section, we will explain the modules that used in *CPH-TS*: Imputing missing data, and acquisition function that selects the most salient regions for profiling, and stopping criteria for TS-A selection. The detailed relationship of these modules are illustrated in Fig. 3

5.1 Imputing Missing Data

Our imputing method Compressive Population Health (CPH), which is based on Generative Adversarial Imputation Nets (GAIN) [44]. Unlike the traditional deep learning based data imputation models, CPH is based on Generative Adversarial Neural-network (GAN) [15], which automatically generates a large amount of synthetic data points for data augmentation in order to enhance the training efficacy. Furthermore, compared with the traditional GAN methods that only use the intra-disease correlation, CPH exploits both intra-disease and inter-disease correlations to impute missing data with a higher accuracy.

Firstly, we input all the collected disease matrices C, and the missing values are replaced with zero, which are fed into CNN [1] for feature engineering to produce the feature matrix, X'. Then we calculate the corresponding mask matrix T with respect to the selected target disease. Lastly, we feed mask matrix (in addition to the feature matrix) into the GAN [15] model for training, after which the imputed matrix $\hat{F}$ can be obtained. Each of these elements is given as follows:

5.1.1 CNN-based Representation. The CNN representation aims to extract the intra-disease and interdisease data correlations from multiple diseases. First, the missing-value cells of each disease matrix are initialized with different noise variables, Z, which is the standard input in a Adversarial Network and can be obtained by drawing from a normal distribution. Then, we consider the overall disease matrices as an image with respective channel to each disease, where the row and column of the image correspond to time and space, and the number of channels denotes the number of disease types. Lastly, we feed the image into the CNN-based representation $\hat{C}$ to obtain a feature matrix X'.

$$X' = \hat{C}(C, Z) \tag{6}$$

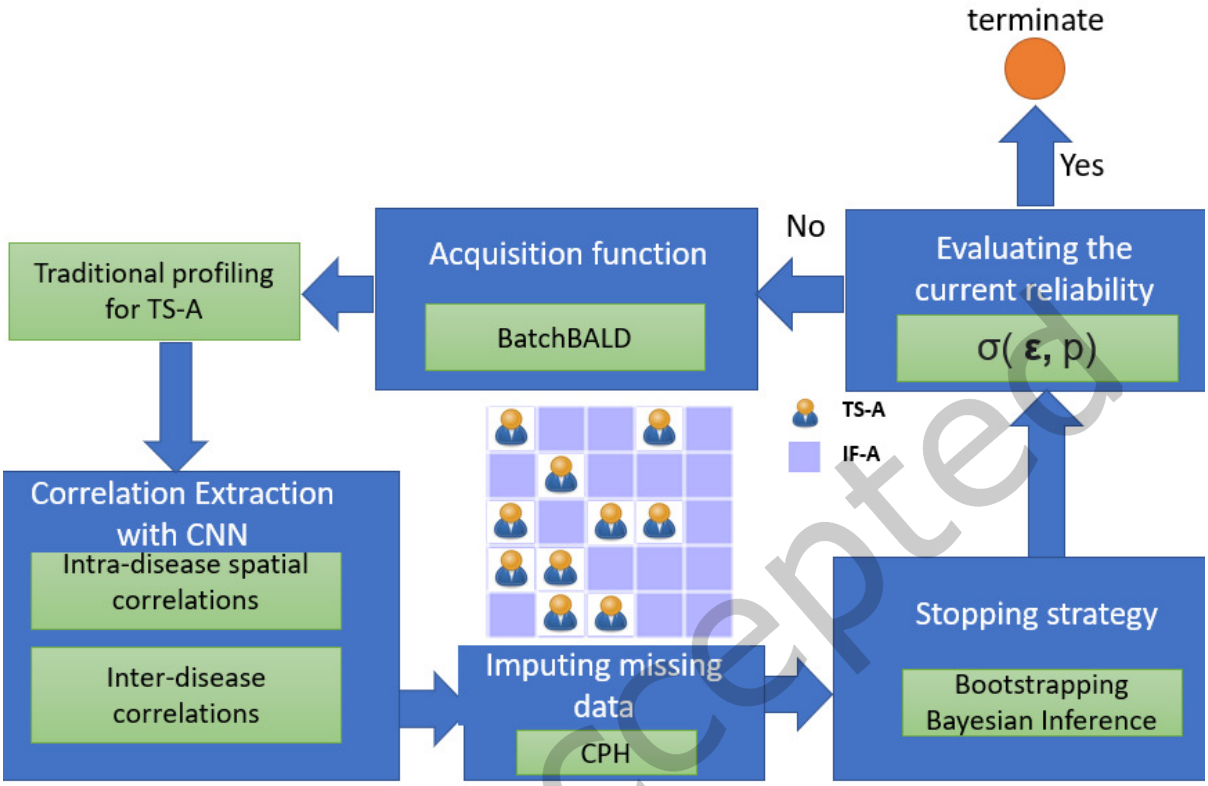

Fig. 3. The detailed relationship of the modules in *CPH-TS*

Notice that wards boundaries dataset provide spatial information of the wards. NCD datasets provide the prevalence rate of each ward between 2008 and 2017. Therefore, these two datasets are mapped in terms of the ward id. Regions are derived from the wards number in London and the cycles are defined by year from 2008 to the current year. In addition, regions are ranked in terms of ward id (where spatially adjacent wards are put together), cycles are ranked by years.

5.1.2 Generator. In the generator network G, feature matrix X' is the input layer and $\hat{F}$ is the output layer, which is a complete matrix. The mask matrix T should change according to the target disease. For example, if diabetes is selected as the target disease, then the mask matrix will tell which cells of the diabetes disease matrix are observed. Let $G : X' \times \{0,1\}^{r \times n} \rightarrow \hat{F}$ be a function, then the matrix $\bar{F}$ and $\hat{F}$ can be given as:

$$\begin{aligned} \bar{F} &= G\left(X', M\right) \\ \hat{F} &= T \odot C + (1 - T) \odot X' \end{aligned} \tag{7}$$

where $\odot$ denotes pair-wise multiplication. C is the collected disease matrix. $\bar{F}$ is the matrix of imputed values, where the value of each cell will be changed in the training process of G, even if the cell is observed. $\hat{F}$ represents the completed data matrix, in which the observed values are taken from the collected disease matrix C and the

missing values are taken from the respective values in the $\bar{F}$. This is a standard GAN setup, where the $\hat{F}$ is the desired output from the training.

5.1.3 Discriminator. In the standard GAN setup, the discriminator is a adversary network to train G. The output of the generator will be classified as either real or fake. However, in our model, the output is a matrix consisted of both observed and imputed values. So, the discriminator will not aim to identify whether the generator output is real or fake, but try to differentiate whether each cell in the matrix is observed or imputed. This is essentially predicting the mask matrix T. The discriminator, therefore, can be given as a function of $D : \hat{F} \rightarrow \{0, 1\}^{r \times n}$, where each cell from the output of D denotes the probability that the corresponding cell of $\hat{F}$ is observed. Therefore, the high probability here indicates that it is likely to be an observed cell.

5.2 Acquisition Function

Acquisition function aims to model the uncertainty from the deep learning prediction. We adopt Batch Bayesian Active Learning by Disagreement (BatchBALD) [23], which is arguably the best acquisition function for the applications of deep active learning until early 2022. It employs the mutual information to estimate the uncertainty between the deep learning predictions and the model parameters. Simply put, it is based on the intuition that obtaining the true label of each cell with a high mutual information can also reveal the hidden model parameters. Let's first introduce the definition of BALD [20], which is the basic unit of BatchBALD:

$$\mathbb{I}\left(y; \omega \mid \boldsymbol{x}, D_{\text{train}}\right) = \mathbb{H}\left(y \mid \boldsymbol{x}, D_{\text{train}}\right) - \mathbb{E}_{\mathrm{p}(\omega \mid D_{\text{train}})}\left[\mathbb{H}\left(y \mid \boldsymbol{x}, \boldsymbol{\omega}, D_{\text{train}}\right)\right] \tag{8}$$

where unlabelled dataset, D_{pool} and D_{train} are the unlabelled dataset and the labelled training dataset respectively. x is a single cell of D_{pool}, y is the true label of x. $p(\omega, D_{train})$ is a Bayesian model over the D_{train} with parameters ω. There are two terms in equation 8, where we want to see the mutual information as high as possible. Therefore, we need to maximize the left term and minimize the right one. The left term denotes the entropy with respect to the model prediction, which will be high when the prediction is uncertain. The right term corresponds to the entropy of the prediction over the posterior distribution of the model parameters. It will be low if the sampling of the model parameters are consistent with the posterior distribution. Hence, for both terms, the best scenarios happen when the data can be explained in many ways, implying that there is a high disagreement among the posterior draws.

Then BatchBALD [23] is built on top of BALD [20], where cells are jointly scored by calculating the mutual information between a set of cells and the model parameters.

$$a_{\text{BatchBALD}}\left(\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_b\}, \mathrm{p}\left(\boldsymbol{\omega} \mid D_{\text{train}}\right)\right) = \mathbb{I}\left(y_1, \ldots, y_b; \boldsymbol{\omega} \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_b, D_{\text{train}}\right) \tag{9}$$

where one can obtain the optimal batch by employing a batch-based greedy algorithm [23] that boils down the problem to selecting the top b highest-scoring cells.

5.3 Stopping strategy

In *CPH-TS*, for each profiling cycle, we need to find the optimal timing to stop TS-A task selection. But the problem is as follows: if we stop prematurely, the data might not be enough to meet the predefined $\sigma(\epsilon, p)$ reliability for the CPH task; On the other hand, if too late, then there will be redundant data, which would incur some additional cost. A good stopping criteria should strike the optimal balance between the data reliability and data redundancy. To achieve this, we propose a Bootstrapping Bayesian-Inference (BBI) method to learn the stopping criterion for each profiling cycle. First, BBI uses leave-one-out bootstrapping re-sampling method [10] to get a set of re-deducted profiling data with the respective ground-truth obtained from the TS-A data. Then, the

re-deduced data is compared to the ground-truth, where Bayesian inference is employed to evaluate if the current data reliability can meet the predefiend $\sigma(\epsilon, p)$ reliability requirement or not (c.f., Section 5.4).

In machine learning, leave-one-out bootstrap is a popular re-sampling method to evaluate the model's performance. Given a training dataset of m observations, the core idea of bootstrapping re-sampling [10] is for each time, we remove one data point and use the remaining $m-1$ data point as training data to re-deduce it. We repeat this process for m times, and get the performance of m predictions along with the respective true observations, which is then used to measure the prediction error.

In each cycle, BBI will remove one data point from the collected dataset and then run the imputation algorithm R to re-deduce it. Once the algorithm finish enumerating the entire collected dataset, we obtain two embeddings x and y, where x denotes the ground-truth for the current cycle, while y is the respective re-deduced data by using leave-one-out bootstrap. Let there be m' collected data from m' regions for each cycle, then both x and y are consisted of m' elements:

$$\mathbf{x} = \langle x_1, x_2, \cdots, x_{m'} \rangle, \quad \mathbf{y} = \langle y_1, y_2, \cdots, y_{m'} \rangle \tag{10}$$

where x_i denote the value of the ith collected region in cycle k, and y_i is the re-deduced value by removing x_i from the collected dataset. Given x and y, we will illustrate how to evaluate whether the task could satisfy the predefined $\sigma(\epsilon, p)$ reliability or not.

5.4 Evaluating the Current Reliability

Formally speaking, if we want to ensure that a task meets the $\sigma(\epsilon, p)$ reliability, we need to make sure that at least p of all data points are in the error bound ϵ.

Hence, we convert the problem of evaluating whether the TS-A task meets the $\sigma(\epsilon, p)$ reliability to calculate $P(\mathcal{E}_k < \epsilon)$, where $\mathcal{E}_k$ is the overall profiling error at cycle k, which can be estimated by using Bayesian inference [3].

If we define $\mathcal{E}_k$ to be an unknown parameter and $g(\mathcal{E}_k)$ is the corresponding prior distribution. Then, after applying acquisition function, we can update the distribution of $\mathcal{E}_k$ in terms of observation θ, which leads to the posterior distribution $g(\mathcal{E}_k|\theta)$, we update the probability distribution as follows (Bayesian Rule [45])

$$g\left(\mathcal{E}_k \mid \boldsymbol{\theta}\right) = \frac{f\left(\boldsymbol{\theta} \mid \mathcal{E}_k\right) g\left(\mathcal{E}_k\right)}{\int_{-\infty}^{\infty} f\left(\boldsymbol{\theta} \mid \mathcal{E}_k\right) g\left(\mathcal{E}_k\right) d\mathcal{E}_k} \tag{11}$$

where $f(\theta|\mathcal{E}_k)$ is the probability of observing θ conditioned on $\mathcal{E}_k$. We can now simplify the problem of estimating $P(\mathcal{E}_k < \epsilon)$ as calculating the posterior distribution $g\left(\mathcal{E}_k \mid \boldsymbol{\theta}\right)$.

$$P\left(\mathcal{E}_k \leq \epsilon\right) \approx \int_{-\infty}^{\epsilon} g\left(\mathcal{E}_k \mid \boldsymbol{\theta}\right) d\mathcal{E}_k \tag{12}$$

If $P(\mathcal{E}_k < \epsilon) > p$, then *CPH-TS* ends the task selection process of the cycle k and the next cycle will be started at due time. Otherwise, *CPH-TS* keeps selecting a batch of new regions to collect profiling data. We use mean absolute error (MAE) as the observation variable θ, which can be estimated with maximum likelihood estimation. We then calculate the posterior of $g(\mathcal{E}_k \mid \boldsymbol{\theta})$ and determine whether more TS-A should be selected.

$$\begin{aligned} \theta &= MAE \\ &= \frac{\sum_i |y_i - x_i|}{m'} \end{aligned} \tag{13}$$

Until now, we need to evaluate the reliability after assigning a new TS-A task, wait for the health expert's results, and then retrain the entire model. This is a time-consuming process, as will be shown in Section 6.7. Can

we save the time cost at certain stage of this process? The answer is Yes. Theoretically, a low-rank matrix M can only be imputed perfectly by using convex optimization [5] when the minimum number of observed entries m in the M satisfy the condition of Eq. 14:

$$m \geq Cn^{1.2} r \log n \tag{14}$$

where n is the maximum dimension of M. In our case n is mostly the number of regions rather than the number of cycles, since the former is normally greater than the latter. C is a positive constant and r is the rank of M, which can be approximated as the number of the cycles. Since the matrices for the **PPP** tasks are not low-rank, the required number of samples should be much larger than m. As will be shown in Sec. 6.7, *CPH-TS* cannot achieve an satisfactory result until the completion of the data collection from the first 10% TS-A regions.

On the other hand, empirically speaking, uncertainty-based active learning approaches initially perform worse than random sampling in a wide range of applications [23, 27, 36, 42], but after a certain number of training acquisition steps (normally between 8% and 20% of training dataset) they begin to improve and supersede the uniform sampling. The above analysis leads us to apply a random sampling approach for all the active learning approaches when assigning the first 10% of the training dataset. Therefore, the imputation algorithm, acquisition function, model retraining, and the stopping strategy can all be omitted to significantly reduce the computational cost and waiting time at the beginning, since it is impossible for the active learning models to attain satisfactory performance and meet the error bound during this stage.

6 EXPERIMENTAL RESULTS

In this section, we use two real-world datasets, National Health Service (NHS), to evaluate *CPH-TS*. We use two publicly available datasets: Ward Boundaries of London [5] and NCDs Prevalence [6]. The former is provided by the UK's mapping bureau with what might be the most accurate geographical statistics given to the public. The collection contains information about 630 London wards, including their unique names, forms, and codes. Between 01/04/2008 and 31/03/2017, the latter was downloaded from the National Health Service, and it comprises three types of NCDs: obesity, diabetes, and hypertension. Each kind is represented by a percentage of all people on the practise roster.

6.1 Experimental Setup

We employ historical training data spanning the years 2008 to 2013 and test data spanning the years 2014 to 2017. For instance, if we change the year to 2014, we can evaluate active learning's performance in 2014 utilizing data from 2008 to 2013. If we use 2016 as the current year, we may evaluate 2016 performance using data ranging from 2008 to 2015. In addition to MAE, we also use RMSE [7] [6] to evaluate the experimental results.

To compare *CPH-TS* with other active learning methods, we use the following baselines:

- FIX-k: An alternative approach that extends the active learning to the TS-A allocation task by fixing the total task number k in each profiling cycle, while still using BALD [23] and DEAL [17] to actively select regions as TS-A; we call this customized algorithm BALD-FIX-k and DEAL-FIX-k respectively. Compared to FIX-k approaches, *CPH-TS* exhibits the power brought by BBI that dynamically determines the best time to stop the TS-A selection task, which in turn can adaptively change the number of task allocation in each cycle.

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

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

[7] https://en.wikipedia.org/wiki/Root-mean-square_deviation

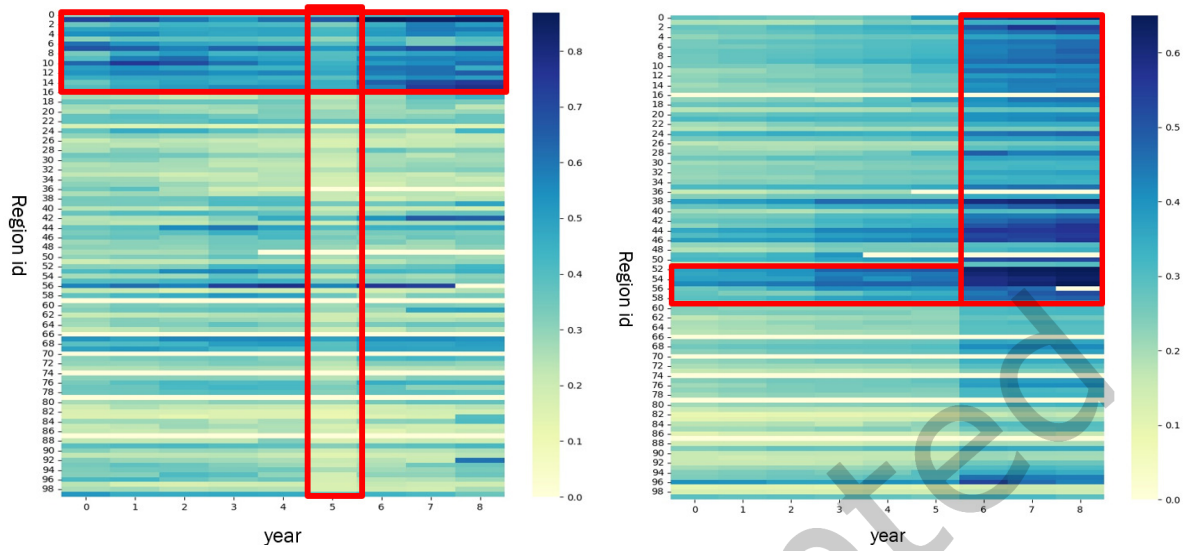

Fig. 4. The temporal-spatial correlations from the input disease matrix (left) and the CNN feature map (right) for obesity.

- RANDOM-TS: An intuitive approach that randomly select the next region for profiling, but still exploits BBI as the stopping criterion. Compared to RANDOM-TA, *CPH-TS* has the benefit of applying BALD to select the salient regions for profiling.

6.2 CNN Representation Analysis

In section 5.1, we enter the global disease matrices as an image, where time and space correspond to the image's row and column. Given that CNN [1] can extract local information from images, it is worthwhile to investigate how the values of disease matrices are influenced by their neighbouring values (time and space). This section demonstrates the usefulness of modelling disease matrices as image data and extracting features using CNN.

In Fig. 4, we highlight the temporal-spatial patterns of the first 100 regions (ranked in terms of ward id) between 2008 and 2016. Left is the disease matrix and right is the last CNN layer for obesity. The last CNN layer is the so-called feature map. From the figure on the left, it is evident that there are temporal spatial patterns. For instance, neighbouring regions are more likely to share a comparable prevalence rate for the first fifteen regions and the fifth year (c.f., the red box in the left figure). It is also noteworthy that the CNN output is capable of capturing a variety of characteristics that are difficult for humans to discern from the CNN input. For instance, the first red box is only marginally correlated in the disease matrix, however the sixth, seventh, and eighth years are highly correlated in the feature map following the CNN transformation process.

6.3 Performance Analysis: Imputing Missing Values

First, we need to evaluate the effectiveness of CPH for imputing missing values, which is then compared to the other state-of-the-art matrix completion algorithms, including NMF [11], ST-KNN [7], Deep Multimodal Encoding (DME) [26], Auto-Encoder [35], and GAIN [44]. Note that for GAIN and CPH, the optimization parameters are identical to that in [12], which has been explained in the code of this paper [8], where the training epoch is 10, learning rate is 0.01, the batch size is 483, which denotes the number of samples in the mini-batch. For NMF,

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

the embedding dimension is 5, the default learning rate is 0.01, and the convergence rate is set as $1e-3$, which controls the stopping training point. The Auto-encoder is based on a MLP structure with four fully connected hidden layers with *relu* as activation function, The initial learning rate is $1e-3$, and the number of training iteration is $1e5$. For ST-KNN, the optimal k value is 6. For DME, we employ the auto-encoder to predict the missing values of each disease.

Methods	2016				2017			
	m= 90%		m= 70%		m=90%		m=70%	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
NMF	0.1518	0.1180	0.1346	0.1064	0.1661	0.1331	0.1513	0.1208
Auto-Encoder	0.0857	0.0616	0.0817	0.0597	0.0772	0.0575	0.0681	0.0520
ST-KNN	0.0794	0.0557	0.0752	0.0546	0.0739	0.0520	0.0632	0.0472
DME	0.0691	0.0525	0.0619	0.0444	0.0694	0.0634	0.0624	0.0459
GAIN	0.0948	0.0597	0.0616	0.0509	0.0617	0.0491	0.0507	0.0415
CPH	**0.0577***	**0.0422***	**0.0455***	**0.0355***	**0.0529***	**0.0410***	**0.0405***	**0.0312***

Table 1. Inference reliability of obesity (* denotes p < 0.05 in the respective t test)

Table 5 shows the overall profiling error of different imputation algorithms on the NHS datasets with varying settings. Notice that we only analyse the results of obesity in this section due to the space limitation. The other two types of diseases actually show even better results, since obesity has the weakest correlation with the other two diseases, therefore its performance is the worst. In this experiment, we consider each profiling cycle k as the latest cycle (we use 2016 and 2017 as the test dataset), imputing the full disease matrix with the collected profiling matrix, and then estimate the overall profiling error in terms of RMSE and MAE [6]. m parameter indicates the percentage of missing data in the training dataset. Consistent with the [12], our evaluation results demonstrate that CPH outperforms other approaches with respect to both metrics with a performance gain ranging from 9.1% to 14.8%, indicating that it is extremely effective at imputing missing health data. It is clear that the traditional NMF method cannot impute the missing values well since it underfits the training dataset. To further evaluate the efficacy of CPH, a two-sample t test is conducted between CPH and GAIN, the best baseline method. Each model is executed 20 times with noise variables Z taken from a normal distribution. The results reveal that the improvement in CPH performance is statistically significant at the 95% confidence level ($p<0.05$). Compared to the deep learning models i.e., Auto-Encoder and GAIN, CPH exploits both inter-disease and intra-disease correlations to infer the unknown values whereas other baselines only utilize the inter-disease correlations. In addition, both auto-encoder and GAIN are essentially unsupervised learning techniques, where different data representations are learned during model training process that can cause divergence to the original data distribution. But in the CPH model, we ensure that the generator output respect the true data distributions by introducing a hinting mechanism. As a result, CPH is used to impute the missing values for the rest of the paper.

6.4 Performance Analysis: Acquisition function

Acquisition function will directly determine the most salient TS-A regions. As a result, it is critical to analyse and find out which acquisition function is the best appropriate for this particular scenario. We compare the performance of several state-of-the-art acquisition functions, namely, BALD sampling [36], Entropy-based sampling [18], and BatchBALD [23] (batch_size = 5). The random sample is also added, which shuffles all the regions and select them one by one at random, to examine what the outcomes would be if no acquisition function is utilised.

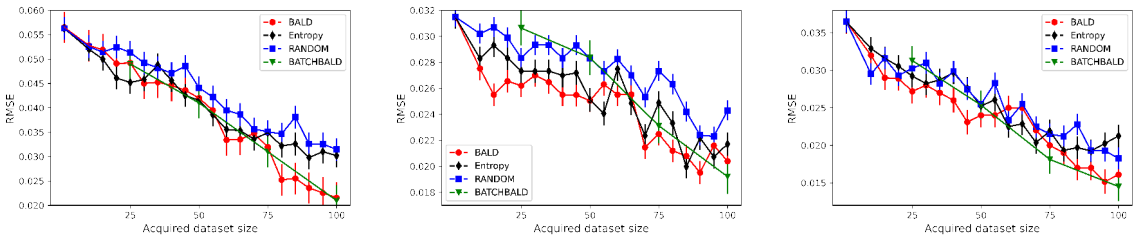

Fig. 5. The performance of varying active learning methods for obesity (left), diabetes (middle) and hypertension(right)

Figure 5 shows the performance comparison of varying active learning methods. Obesity has the best performance boost when active learning is applied. This is because it has the largest imputation error at the beginning, implying that there is significant room for improvement. Also, as expected, adding additional TS-A regions can generally improve the performance for all the methods across three diseases. Entropy methods shows a similar performance to that of random sampling, perhaps this approach is not very effective for this dataset with many repeated values. BatchBALD shows comparable performance than BALD, even though it selects a batch of regions in each iteration. This is probably because unlike BALD, which solely takes uncertainty into account, BatchBALD integrates correlation between samples by calculating the batch's mutual information between the sample embeddings and model parameters. In general, however, BALD-based methods exhibits superior results compared to the other two, as they take account both the prediction uncertainty and the posterior of the model parameter by exploiting Bayesian inference. In practise, rather than acquiring individual data points with each acquisition step, batches of data points are obtained to minimise the model-retrained time and expert-time. Model retraining is a computational constraint for larger models where the expert's time is usually expensive: consider the effort required to commission a medical expert to investigate a single region, pause there for the model to be retrained, and then commission a new medical expert to investigate the next TS-A region, as well as the additional time required. With its competitive performance and reduced computation time, BatchBALD is the natural choice for *CPH-TS*.

6.5 Performance Analysis: Stopping Strategy

To evaluate the effectiveness of the proposed stopping strategy, we need to have several disparage settings of $\sigma(\epsilon, p)$ to see what percentage of profiling cycles would keep the overall profiling error less than the predefined error bound ϵ. For p, we set it with a large value, namely 0.90 and 0.95, i.e., which ensures that most (90% or 95%) profiling cycles' error will be less than ϵ, which is a principled and realistic setting for most of the *PPP* scenarios. For ϵ, we set it between 5% and 10%.

	obesity		diabetes		hypertension	
ϵ	5%	10%	5%	10%	5%	10%
$p = 0.9$	0.904	0.91	0.912	0.917	0.901	0.912
$p = 0.95$	0.937	0.973	0.958	0.963	0.927	0.955

Table 2. The percentage of cycles whose error are smaller than the error bound

For any given error bound, Table 2 shows that the actual percentage of the cycles whose errors are smaller than the error bound ϵ, is quite close to the p set in the predefined $\sigma(\epsilon, p)$ requirement. For instance, for $p = 0.90$,

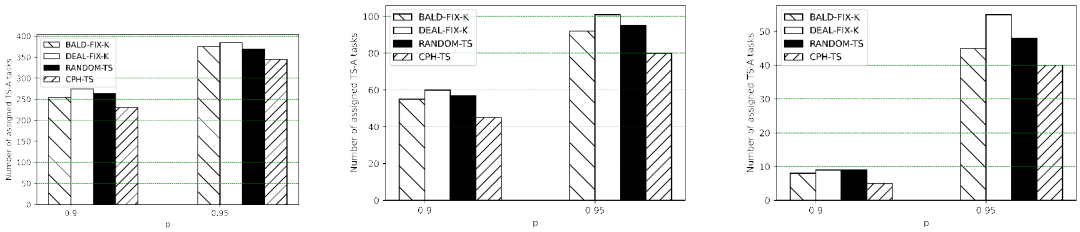

Fig. 6. Number of assigned tasks of obesity (left), diabetes (middle), and hypertension (right) respectively.(ϵ = 5% , batch size = 5, varying p).

all the actual percentage are larger than 90%; for $p = 0.95$, while some actual values are slightly less than 95%, they are all quite close to 95% (in our settings, the worst actual percentage is 0.927, which is only about 0.023 smaller than 0.95). Based on these results, we confirm that by using BBI as the stopping strategy, *CPH-TS* can well meet the predefined $\sigma(\epsilon, p)$ requirement of reliability.

6.6 Performance Analysis: Cost

The cost of the project is largely determined by the number of assigned tasks of **PPP**. Fig. 6 illustrates the minimum number of assigned tasks for varying methods when the stopping criteria is satisfied. For hypertension and $p = 0.9$, *CPH-TS* only needs to assign 5 TS-A tasks, whereas BALD-FIX-K and DEAL-FIX-K requires to assign 8 and 9 TS-A tasks respectively; for hypertension and $p = 0.95$, *CPH-TS* assign forty tasks while RAND-TS, DEAL-FIX-K and BALD-FIX-K need to assign 48, 55, and 45 tasks respectively. 5 tasks seems to be a small number, but relatively speaking it allows us to reduce the cost by 16.7%, 27.3% and 11.1% respectively. We can see a similar pattern on the results of diabetes. Obesity, however, seems to have a disparage pattern compared with the other two. Specifically, *CPH-TS* assigns 38% (56.9%) regions to ensure the minimum overall reliability criteria in 90%(95%) of the cycles, which is somewhat an abnormal performance. After a careful examination of the dataset, we find that the Pearson correlation between obesity and diabetes is 0.58, and correlation between obesity and hypertension is 0.51, which is a relative low value compared with the correlation between diabetes and hypertension. This finding is reported in Fig. 7, which implies that *CPH-TS* works well only if there exist strong correlation among diseases. However, if this assumption cannot hold then *CPH-TS* will require more resources.

In summary, despite *CPH-TS*'s limited efficacy against obesity, it can nevertheless accomplish impressive results in general. It assigns tasks to an average of 21.67% (37.7%) of regions, while ensuring that the overall profiling error is less than 0.05 in 90% (95%) of cycles. Even in the worst-case scenario of obesity, *CPH-TS* can still save 43.1% of the cost of expert-led research, which we argue will be a significant workload.

6.7 Performance Analysis: Computational Time

	obesity		diabetes		hypertension	
ϵ	5%	10%	5%	10%	5%	10%
p=0.9	6.2h	4.5h	1.9h	0.9h	2.4h	1.5h
p=0.95	8.5h	5.7h	2.5h	2.3h	3.3h	2.8h

Table 3. Computational cost under varying settings

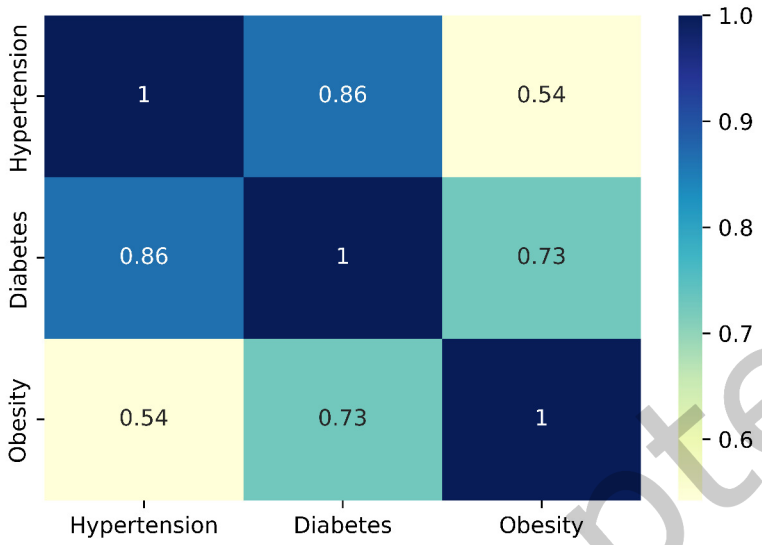

Fig. 7. The correlation among varying NCDs

For real-world application, we also need to know the computation time of *CPH-TS* to check if it can meet the practical **PPP** scenario. The experiments are conducted on a workstation (Intel core i9-10900k, 2 × 3080 GPU, 128GB RAM, Windows 10) with Python 3.7. Table 3 reports the computation time for different diseases of *CPH-TS*. The most computation intensive part is the imputation algorithm, CPH, which is a deep learning architecture and thus needs 8.5 hours in the worse case (when the disease is obesity with $p = 0.95, \epsilon = 5\%$). Note that we skipped the reliability check for the first m collected TS-A. The reason is already explained in Section 5.3, Eq. 14. Similar to [32], we set $C = 1$ and $r = k$, where k is the number of the cycle. This operation enables for the *CPH-TS* to skip the reliability check of the first 60 TS-A tasks in each cycle. In addition, it is possible to apply CPH to a parallelized setting to further reduce the computation time.

In a nutshell, *CPH-TS* spends on average 18 minutes to complete one iteration of data imputation and another 10 seconds to assign new tasks, i.e., evaluating the overall data reliability and, if it cannot meet the predefined requirement, then finds the next profiling region. Therefore, if we assume that no time is needed for health expert to collect the profiling data, *CPH-TS* can assign tasks to 28 regions (when $batchsize = 5$) for one disease in one day in the worse scenario; when the health experts need some time to complete the data collection, we can obtain the new task assignments from *CPH-TS* overnight and dispatch the new TS-A tasks the next day. We believe it can tackle most real-world profiling problems with high efficiency.

7 CONCLUSION AND LIMITATIONS

We aim to reduce the number of required profiling regions and consequently the number of tasks assigned to participants in **PPP** activities in this research. Towards that end, we propose the *CPH-TS* framework, which combines state-of-the-art compressive population health, data augmentation, and active learning mechanisms to

actively select a minimum number of profiling regions in each cycle while imputing missing values for remaining regions and ensuring that the overall data reliability satisfy a predefined error bound.

While the experiments on a real-world NHS dataset demonstrate that *CPH-TS* is capable of learning intra-disease and inter-disease data correlations from previous data and then using them to select minimum number of salient regions for completing the current year's prevalence rate, we have not yet explored how to make *CPH-TS* sustainable year after year. As time passes, historical data will comprise both collected and inferred data entries, so there is a possibility that the inaccuracy in the imputed data entries will be propagated and compounded upon the current year's adoption of *CPH-TS*. As a result, developing a sustainable *CPH-TS* is an exciting and demanding direction for the future. One possibility is to quantify the accuracy of imputed entries, which will be used as weights in the model of prevalence inference. Before CPH is implemented in a new year, the weights will be dynamically modified as even more ground-truth data becomes available, which in turn would further improve the performance of *CPH-TS*.

Moreover, we intend to improve this work by reassessing some of the study's assumptions. For instance, in practise, the assumption of abundant Participants may not necessarily hold true for all **PPP** applications. It is possible that the organiser will be unable to locate an expert to perform a task in the designated salient region for some cycles. In that circumstance, the work allocation problem cannot be boiled down to an region selection problem, and therefore we must handle the issue by taking both the participants' availability and the expert's quality into account.

Lastly, CPH-TS essentially selects TS-A regions for each disease separately, which is inefficient for profiling. In the future work, we may consider multi-task deep active learning.

8 ACKNOWLEDGEMENTS

This work was supported by EPSRC New Investigator Award under Grant No EP/V043544/1.

REFERENCES

[1] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. 2017. Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)*. Ieee, 1–6.

[2] Jannah Baker, Nicole White, and Kerrie Mengersen. 2014. Missing in space: an evaluation of imputation methods for missing data in spatial analysis of risk factors for type II diabetes. *International journal of health geographics* 13, 1 (2014), 1–13.

[3] George EP Box and George C Tiao. 2011. *Bayesian inference in statistical analysis*. John Wiley & Sons.

[4] Robert Burbidge, Jem J Rowland, and Ross D King. 2007. Active learning for regression based on query by committee. In *International conference on intelligent data engineering and automated learning*. Springer, 209–218.

[5] Emmanuel J Candès and Benjamin Recht. 2009. Exact matrix completion via convex optimization. *Foundations of Computational mathematics* 9, 6 (2009), 717–772.

[6] Tianfeng Chai and Roland R Draxler. 2014. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific model development* 7, 3 (2014), 1247–1250.

[7] Shifen Cheng, Feng Lu, Peng Peng, and Sheng Wu. 2018. Short-term traffic forecasting: an adaptive ST-KNN model that considers spatial heterogeneity. *Computers, Environment and Urban Systems* 71 (2018), 186–198.

[8] Jae Won Cho, Dong-Jin Kim, Yunjae Jung, and In So Kweon. 2022. Mcdal: Maximum classifier discrepancy for active learning. *IEEE transactions on neural networks and learning systems* (2022).

[9] Archie CA Clements, Marie-Alice Deville, Onésime Ndayishimiye, Simon Brooker, and Alan Fenwick. 2010. Spatial co-distribution of neglected tropical diseases in the East African Great Lakes region: revisiting the justification for integrated control. *Tropical medicine & international health* 15, 2 (2010), 198–207.

[10] Bradley Efron. 2012. Bayesian inference and the parametric bootstrap. *The annals of applied statistics* 6, 4 (2012), 1971.

[11] Julian Eggert and Edgar Korner. 2004. Sparse coding and NMF. In *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)*, Vol. 4. IEEE, 2529–2533.

[12] 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*. 183–193.

[13] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*. PMLR, 1050–1059.
[14] Caroline W Gitonga, Peris N Karanja, Jimmy Kihara, Mariam Mwanje, Elizabeth Juma, Robert W Snow, Abdisalan M Noor, and Simon Brooker. 2010. Implementing school malaria surveys in Kenya: towards a national surveillance system. *Malaria journal* 9, 1 (2010), 1–13.
[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).
[16] Sangeeta Grover and Gagangeet Singh Aujla. 2015. Twitter data based prediction model for influenza epidemic. In *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE, 873–879.
[17] Patrick Hemmer, Niklas Kühl, and Jakob Schöffer. 2022. Deal: deep evidential active learning for image classification. In *Deep Learning Applications, Volume 3*. Springer, 171–192.
[18] Alex Holub, Pietro Perona, and Michael C Burl. 2008. Entropy-based active learning for object recognition. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 1–8.
[19] Roger A Horn. 1990. The hadamard product. In *Proc. Symp. Appl. Math*, Vol. 40. 87–169.
[20] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745* (2011).
[21] David Janz, Jos van der Westhuizen, and José Miguel Hernández-Lobato. 2017. Actively learning what makes a discrete sequence valid. *arXiv preprint arXiv:1708.04465* (2017).
[22] Linli Jiang, Chao-Xiong Chen, and Chao Chen. 2023. L2mm: learning to map matching with deep models for low-quality gps trajectory data. *ACM Transactions on Knowledge Discovery from Data* 17, 3 (2023), 1–25.
[23] Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. 2019. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems* 32 (2019), 7026–7037.
[24] Mareike Kroll, Revati K Phalkey, and Frauke Kraas. 2015. Challenges to the surveillance of non-communicable diseases–a review of selected approaches. *BMC public health* 15, 1 (2015), 1–12.
[25] Jun Li, Juliane Manitz, Enrico Bertuzzo, and Eric D Kolaczyk. 2021. Sensor-based localization of epidemic sources on human mobility networks. *PLoS computational biology* 17, 1 (2021), e1008545.
[26] Zuozhu Liu, Wenyu Zhang, Shaowei Lin, and Tony QS Quek. 2017. Heterogeneous sensor data fusion by deep multimodal encoding. *IEEE Journal of Selected Topics in Signal Processing* 11, 3 (2017), 479–491.
[27] Markus Lucero. 2021. Evaluating the Effectiveness of Active Learning Methods in Predicting Biochemical Properties.
[28] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 825–832.
[29] Liantao Ma, Xinyu Ma, Junyi Gao, Xianfeng Jiao, Zhihao Yu, Chaohe Zhang, Wenjie Ruan, Yasha Wang, Wen Tang, and Jiangtao Wang. 2021. Distilling Knowledge from Publicly Available Online EMR Data to Emerging Epidemic for Prognosis. In *Proceedings of the Web Conference 2021*. 3558–3568.
[30] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 833–840.
[31] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. 2021. A survey of deep active learning. *ACM Computing Surveys (CSUR)* 54, 9 (2021), 1–40.
[32] Yingying Ren, Yuxin Liu, Ning Zhang, Anfeng Liu, Neal N Xiong, and Zhiping Cai. 2018. Minimum-cost mobile crowdsourcing with QoS guarantee using matrix completion technique. *Pervasive and Mobile Computing* 49 (2018), 23–44.
[33] Donald B Rubin. 2004. *Multiple imputation for nonresponse in surveys*. Vol. 81. John Wiley & Sons.
[34] Annie Sauer, Robert B Gramacy, and David Higdon. 2022. Active learning for deep Gaussian process surrogates. *Technometrics* (2022), 1–15.
[35] Meng Shen, Huaizheng Zhang, Yixin Cao, Fan Yang, and Yonggang Wen. 2021. Missing Data Imputation for Solar Yield Prediction using Temporal Multi-Modal Variational Auto-Encoder. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2558–2566.
[36] Toan Tran, Thanh-Toan Do, Ian Reid, and Gustavo Carneiro. 2019. Bayesian generative active deep learning. In *International Conference on Machine Learning*. PMLR, 6295–6304.
[37] JW Van Groenigen, M Gandah, and J Bouma. 2000. Soil sampling strategies for precision agriculture research under Sahelian conditions. *Soil Science Society of America Journal* 64, 5 (2000), 1674–1680.
[38] Richard G Wamai, Andre Pascal Kengne, and Naomi Levitt. 2018. Non-communicable diseases surveillance: overview of magnitude and determinants in Kenya from STEPwise approach survey of 2015. *BMC Public Health* 18, 3 (2018), 1–8.
[39] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. 2016. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology* 27, 12 (2016), 2591–2600.

[40] Leye Wang, Daqing Zhang, Animesh Pathak, Chao Chen, Haoyi Xiong, Dingqi Yang, and Yasha Wang. 2015. CCS-TA: Quality-guaranteed online task allocation in compressive crowdsensing. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 683–694.
[41] Yingzi Wang, Xiao Zhou, Cecilia Mascolo, Anastasios Noulas, Xing Xie, and Qi Liu. 2018. Predicting the Spatio-Temporal Evolution of Chronic Diseases in Population with Human Mobility Data. IJCAI.
[42] Yazhou Yang and Marco Loog. 2016. Active learning using uncertainty information. In *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2646–2651.
[43] Donggeun Yoo and In So Kweon. 2019. Learning loss for active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 93–102.
[44] Jinsung Yoon, James Jordon, and Mihaela Schaar. 2018. Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*. PMLR, 5689–5698.
[45] Zijian Zheng and Geoffrey I Webb. 2000. Lazy learning of Bayesian rules. *Machine Learning* 41, 1 (2000), 53–84.
[46] Jia-Jie Zhu and José Bento. 2017. Generative adversarial active learning. *arXiv preprint arXiv:1702.07956* (2017).

9 APPENDIX

Due to the space limitation, we only reported the performance of CPH for obesity in Table 1. The CPH performance comparison on diabetes and hypertension are reported in Table 4 and 5, respectively. The results from these two tables are similar to that of Table 1, namely, the performance gain of CPH over GAIN is statistically significant at 95% confidence interval.

Methods	2016				2017			
	m= 90%		m= 70%		m=90%		m=70%	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
NMF	0.1715	0.1363	0.1519	0.1199	0.2683	0.2234	0.2109	0.1758
Auto-Encoder	0.1228	0.1061	0.1189	0.1022	0.1139	0.0979	0.1101	0.940
ST-KNN	0.1170	0.0997	0.1096	0.0903	0.0878	0.0764	0.0597	0.0508
DME	0.1003	0.0862	0.0790	0.0678	0.0971	0.0834	0.0777	0.0660
GAIN	0.1446	0.1201	0.1054	0.0852	0.01658	0.1449	0.1043	0.0838
CPH	**0.0597***	**0.0239***	**0.0278***	**0.0187***	**0.0310***	**0.0221***	**0.0297***	**0.0212***

Table 4. Inference reliability of diabetes (* denotes p < 0.05 in the respective t test)

Methods	2016				2017			
	m= 90%		m= 70%		m=90%		m=70%	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
NMF	0.0919	0.0719	0.0887	0.0713	0.0987	0.0784	0.0795	0.0626
Auto-Encoder	0.0530	0.0392	0.0524	0.0383	0.0498	0.0379	0.0468	0.0365
ST-KNN	0.0461	0.0344	0.0469	0.0345	0.0363	0.0287	0.0324	0.0255
DME	0.0397	0.0298	0.0352	0.0246	0.0402	0.0310	0.0360	0.0261
GAIN	0.0371	0.0283	0.0241	0.0167	0.0286	0.0205	0.0199	0.0141
CPH	**0.0327***	**0.0261***	**0.0198***	**0.0159***	**0.0230***	**0.0162***	**0.0148***	**0.0112***

Table 5. Inference reliability of hypertension (* denotes p < 0.05 in the respective t test)