Volume 8, Issue 6
December 2019

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

Secrecy Strengthening Concert (Privacy Protection) In Data Mining

[1]T. Pavithra and [2] Dr. K. Thangadurai

[1] PG and Research Department of Computer Science, Research Scholar,
Government Arts College (Autonomous), Karur

[2]PG and Research Department of Computer Science, Assistant Professor & Head,
Government Arts College (Autonomous), Karur

E-mail: [1]pavithrakrishnaveni2011@gmail.com, [2] ktramprasad05@gmail.com

ABSTRACT

Data mining techniques have been broadly used in many research disciplines such as medicine, life sciences, social sciences, etc. to extract useful knowledge from research data in the form of data mining models. Research data often needs to be available along with the data mining model for verification. The privacy of the published data which needs to be protected because otherwise the published data is subject to misuse such as linking attacks. Therefore, using various privacy protection methods becomes essential. Thus the published models cannot be verified using the cleaned data. This paper says about a technique that not only protects privacy, but also guarantees that the same model, in the form of decision trees or regression trees, can be built from the cleaned data.

Keywords: *Data management, experimentation, twitter*

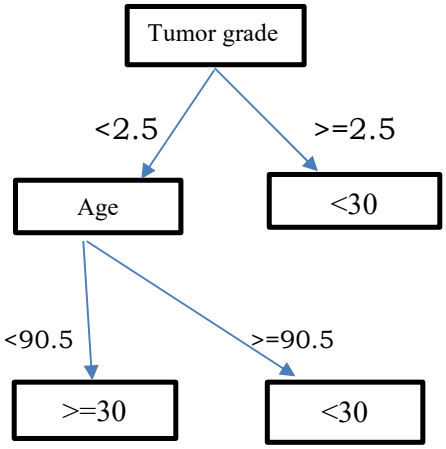
1. INTRODUCTION

Data mining techniques have been commonly used in many research areas such as medicine, life sciences, and social sciences to extract useful information from research data in the form of data mining models. For example, decision trees have been used to predict adverse drug reactions using clinical trial data [1].

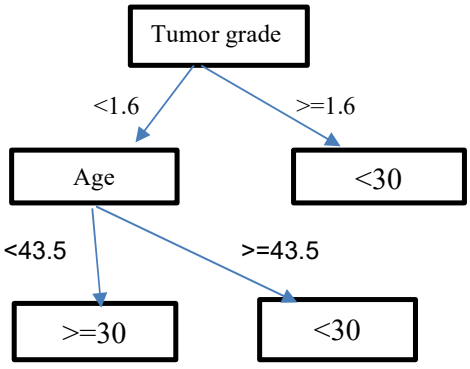
First, research data, if published, can be used by other researchers to verify the published research results. This can significantly add credibility to the results and alleviate some of the problems about scientific misconduct and research fraud. In a survey participated by 1389 researchers in the European Union [2], around 90% of the participants considered that publishing research data was very important or important for validation of research results.

Second, other researchers may conduct secondary analysis over the published research data in their own research. This has been widely used in disciplines such as social science and medical research. In these disciplines, data collection is often very expensive and secondary analysis saves resources that would otherwise be spent on collecting data. For example, secondary analysis was used to discover the causes of some diseases from medical records.

In the European Union survey mentioned before [2], 91% of participants considered that publishing research data was very important or important for reanalysis of existing data. Secondary analysis can be also divided into two categories: (1) reanalysis, which is the analysis of the data on the same research problem and (2) analysis that is used to solve a different research problem.



(a) Original data, accuracy=0.73



(b) Sanitized data, accuracy=0.59

Figure 1 Decision trees built from original cancer data and sanitized data.

Volume 8, Issue 6
December 2019

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

For example, Figure 1(a) shows a decision tree built from a dataset about cancer patients to predict whether a patient will survive more than 30 months. There are three attributes in the dataset: tumor grade, age, and survival status (whether the patient survives more than 30 months). We used a sanitization method proposed in [3]. Figure 1(b) shows the decision tree built from the sanitized data. Thus, other researchers cannot use the sanitized data to verify the published original decision tree.

The decision tree model is not preserved because the decision tree building algorithm needs to computer information gain for all possible splits in the data. Since data values have been distorted during the sanitization process, the information gain computed in the sanitized data is often different from that in the original data. Thus different splits are selected in the sanitized data. For example, in the original data, the split on tumor grade = 2.5 generates the highest information gain. However, after sanitization, the values of tumor grade attribute have been distorted and the best split becomes tumor grade = 1.6.

2. RELATED WORK

Here, first we discuss the risks to privacy and the general categories for the privacy models. Next we discuss privacy protection techniques and briefly mention the work about hiding sensitive patterns. Finally, we place study in the context of the related work.
Privacy risks and models. There are two types of privacy risks.

- Identity disclosure when the identity of a specific person in the dataset is revealed.

- Value disclosure when the values of some sensitive attribute values are revealed.

The two most popular privacy models are K-anonymity and L-diversity. K-anonymity prevents identity disclosure caused by linking attacks, which link attributes such as birth date, gender, and ZIP code with publicly available datasets. This can be done by generalization, that is, replacing specific values with more general ones. For example, the exact age of a patient can be replaced with a range. The same quasi-identifier values form an equivalence class. K-anonymity ensures that there are at least K people with the same quasi-identifier such that the risk of identity disclosure is reduced to 1/K.

L-diversity prevents value disclosure by further requiring that the people with the same quasi-identifier contain at least L well-represented sensitive values such that attackers cannot discover the values of sensitive attributes easily. A more advanced model called t-closeness tries to make sure the distribution of sensitive attributes in each equivalence class is similar to the global distribution [4].

A. Privacy protection techniques

There has been an amusing body of work to enforce privacy protection models [5]. These techniques can be divided into random perturbation [6], generalization, or suppression [7], random permutation (e.g., randomly permute the values of sensitive attributes) [8], and synthetic data generation [9]. There also exists work on secure multiparty computation [10], which is useful for the distributed mining case. In this article we will only consider the case when the research data is published along with mining models. Next we will discuss several techniques related to data publication.

An additive perturbation technique was proposed in [2]. A reconstruction technique was also proposed to reconstruct the marginal distribution from perturbed data. A tree-based approach was proposed to sanitize data [11]. The proposed approach used a KD-tree to divide data into groups and then generalize data in each group. A workload-aware anonymization approach was proposed in [12], where the anonymization process is optimized for specific mining tasks. For example, the anonymization tries to maximize information gain (which is used in decision tree building) for classification. Another perturbation approach was proposed in [11] for categorical data. This approach randomly swaps sensitive attribute values in records that have high disclosure risks and at the same time tries to preserve both the marginal distribution of the sensitive attribute and the correlation between no sensitive attributes and the sensitive attribute.

B. Comparison of our approach with the related work

All the existing work on privacy protection does not guarantee that the decision tree or regression tree models are preserved. The approach proposed in this article preserves these two tree models and at the same time protects data privacy. This article is also an extended version of our preliminary work [12]. The extensions include: (1) more comprehensive experiments, (2) extension of our approach to satisfy given privacy requirements, (3) efficiency improvement of our approach.

3. PROPOSED METHODOLOGY

In this section 3.1, first we specifically formulate the problem tackled here. Then, we briefly describe the decision tree and regression tree building algorithms. In Section 3.2, we prove a theorem that describes the conditions under which a tree model can be preserved. Finally, in Section 3.3, we present a method that preserves both privacy and the decision or regression tree model.

3.1 BACKGROUND

Problem description Let T be a data table with attributes A1, A2…. Am. These attributes can be divided

Volume 8, Issue 6
December 2019

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

into sensitive attributes (whose values need to be protected) and non-sensitive attributes. We also assume that all non-sensitive attributes are quasi-identifier attributes. We assume that attribute Am is the response variable (which needs to be predicted). Let K and L be two integers, and B be a decision tree or regression tree building algorithm. The goal is to create a clean table T_ such that T_ satisfies K-anonymity and L-diversity, and at the same time, B can build the regression tree P from T_ or T to predict the value of Am.

A. Decision tree and regression tree building algorithms

The structure of a decision tree or a regression tree is as follows. Each internal node of a decision tree or regression tree contains a test condition and several branches representing test outcomes. For example, in the root node of the tree seen in Figure 1(a), the patients with a tumor grade less than 2.5 are assigned to the left child, and those with a tumor grade greater than or equal to 2.5 are assigned to the right child. A leaf of a decision tree predicts a class label; a leaf of a regression tree predicts a numerical outcome. These algorithms stop when a certain stopping criterion is met during the successive splitting actions. There are three commonly used splitting criteria for decision trees: information gain, gain ratio, and Gini index. Here we just describe information gain while our approach also applies to the other two.

3.2 PROPOSED DATA SANITIZATION PROCEDURE

Figure 2 describes the Tree-Pattern-Preserving Algorithm (TPP). The input of the algorithm includes original data T, a decision tree or regression tree building algorithm B, and privacy parameters K and L. The output is a tree model P and a sanitized dataset T_ that satisfies both K-anonymity and L-diversity. The same tree P can be built from T_ as well.

Step 1 of the algorithm builds a decision tree with one node. Steps 2 to 4 try to sanitize the data. We will show shortly that these steps satisfy all conditions in Theorem 1 and thus preserve the current decision tree or regression tree. Step 5 will check whether privacy requirements are satisfied. If so, we will repeatedly expand the decision tree or regression tree and rerun steps 2 to 5 to sanitize the data. Otherwise, we return the latest tree that satisfies the privacy requirements along with the sanitized data.

1. Run tree building algorithm B to generate a tree P with only one node.
2. For each attribute A_i that is not the response variable and is not used in P, replace its value with a single value (for categorical attribute, use ALL; for numerical attribute, use mean of A_i).
3. For each numerical attribute A_i that appears in the tree, do the following:
   a) For each node x in P that uses A_i as split attribute, collect boundary values as the maximal A_i value in the left child and the minimal A_i value in the right child.
   b) Sort values of A_i and divide them into intervals using boundary values collected in step 3a)
   c) If an interval contains two boundary values, split it into two equal size intervals such that each contains only one boundary value. Compute the mean of each new interval, let them be $u_1, u_2, ...$
   d) For each node x in P that uses A_i as split attribute, let v_1 (v_2) be the maximal (minimal) A_i value in the left (right) child of x. Let I_1 (I_2) be the intervals with v_1 (v_2) as the right (left) boundary. Compute $d = min\{v_1 - u_1, u_2 - v_2\}$. Replace values in I_1 with v_1-d, and values in I_2 with v_2+d.
4. For each categorical attribute A_i that appears in P, if two-way split is used, divide values of A_i into groups such that the values in the same group appear in the same branches in P (this can be done by sorting values on the branches they appear). Replace values of A_i in the same group with the same generalized value.
5. Group all records on quasi-identifier attributes. For each group, check whether it satisfies K-anonymity and L-diversity.
6. If privacy requirements are satisfied, call tree building algorithm B to expand P once more and rerun step 2 to 5 until the stopping condition of B is met.
7. Otherwise, return the last tree that satisfies the privacy requirements and the data sanitized based on that tree.

Figure 2 Tree-Pattern-Preserving algorithm (TPP).

Next, we show how steps 2 to 4 satisfy conditions in Theorem 1. First, these steps do not change the values of response variables. Thus, Condition (1) is satisfied. Step 4 sanitizes categorical attributes and it is easy to verify it satisfies Condition (2). Step 3 sanitizes numerical attributes. We will use an example to show how it satisfies Condition (3). Figure 3 shows how step 3 works for Example 1. Suppose the tree building algorithm selects a numerical attribute Ai as the split attribute. The best split in original data is between $r3$ and $r4$. Thus step (3a) will pick the Ai values of $r3$ and $r4$ as boundaries (let them be $v1$ and $v2$, respectively). In step (3b), two intervals get created: $I1$ containing $r1$ to $r3$ and $I2$ containing $r4$ to $r6$. Each interval only contains one boundary value. Step (3c) computes the mean of each interval. Step (3d) computes the gap between $v1$ ($v2$) and the mean of $I1$ ($I2$). Let δ be the smaller of these two gaps. It then generalizes the values in $I1$ to $v1 - d$, and values in $I2$ to $v2 + d$. Clearly, the new split value in the sanitized data $(v1 - d + v2 + d)/2$ is the same as the old split value $(v1 + v2)/2$. Thus Condition (3b) is satisfied. The order is also preserved because $v1 \le v2$ and $v1 - d \le v2 + d$. Thus Condition (3a) is satisfied.

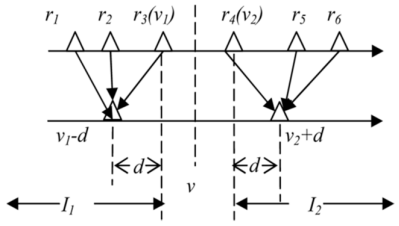


Figure 3 Sanitizing numerical attribute.

4. EXPERIMENTAL EVALUATION

Data: We used two real-life datasets: The Adult dataset from UCI Repository of Machine Learning datasets [13] and the Cancer dataset obtained from

Volume 8, Issue 6
December 2019

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

University of Kentucky Cancer Research Center. The Adult data contains census data and is also the de facto benchmark in the literature. It contains 30717 records, 5 numerical attributes, and 7 categorical attributes. We used "occupation" as the sensitive attribute and the rest as quasi-identifiers. The Cancer dataset contains 3537 records. It has 3 numerical attributes and 3 categorical attributes. We used "histology" as the sensitive attribute. Our method was implemented in R. The experiment was run on a desktop PC with 3.2G HZ CPU and 2GB RAM, running Windows XP.

Methods: For the Adult dataset, we built a decision tree to predict whether the annual household income is over 50K. For the Cancer dataset, we built a regression tree to predict the number of years a patient will survive after diagnosis of cancer. We compare our method (TPP) to the InfoGain method in [3] because it has the best prediction accuracy among existing methods. InfoGain partitions data into groups such that information gain is maximized. It then generalizes quasi-identifier attributes in each group. It does not satisfy Condition 3 in Theorem 1 (preserving order and split values for numerical attributes), thus it does not preserve decision trees or regression trees.

Metric: We reported the accuracy of mining models built from the sanitized data using 10-fold cross-validation. We used K-anonymity and L-diversity to measure the degree of privacy protection. Larger $K$and L mean more protection. In terms of L-diversity, the sensitive attributes in both datasets are not used in the decision tree or regression tree model and are thus suppressed by both TPP and InfoGain. This is the best a privacy protection method can do. The best strategy for attackers is to assume that the sensitive attribute always has the most frequent value, assuming that attackers know the most frequent value of the sensitive attribute.

1. Run tree building algorithm B_1, B_2Ö , B_u to generate top level of trees P_1, P_2, Ö, P_u.
2. For each attribute A_i that is not the response variable and is not used in any of P_1, P_2, … , P_u, replace its value with a single value (for categorical attribute, use ALL; for numerical attribute, use mean of A_i).
3. For each numerical attribute A_i that appears in at least one tree, do the following:
   a) For each node x in P_j that uses A_i as split attribute, collect boundary values as the maximal A_i value in the left child and the minimal A_i value in the right child.
   b) Sort values of A_i and divide them into intervals using boundary values collected in step 3a) from all trees
   c) If an interval contains two boundary values, split it into two equal size intervals such that each contains only one boundary value. Compute the mean of each new interval, let them be u_1, u_2, …
   d) For each node x in P that uses A_i as split attribute, let v_1 (v_2) be the maximal (minimal) A_i value in the left (right) child of x. Let I_1 (I_2) be the intervals with v_1 (v_2) as the right (left) boundary. Compute $d = min\{v_1 - u_1, u_2 - v_2\}$. Replace values in I_1 with v_1-d, and values in I_2 with v_2+d.
4. For each categorical attribute A_i that appears in any of P_1, P_2, Ö, P_u, if two-way split is used, divide values of A_i into groups such that the values in the same group appear in the same branches in P_1, P_2, Ö, P_u (this can be done by sorting values on the branches they appear). Replace values of A_i in the same group with the same generalized value.
5. Group all records on quasi-identifier attributes. For each group, check whether it satisfies K-anonymity and L-diversity.
6. If privacy requirements are satisfied, call tree building algorithm B_1, Ö , B_u to expand all trees once more and rerun step 2 to 5 until the stopping conditions are met.
7. Otherwise, return the last trees that satisfy the privacy requirements and the data sanitized based on these trees.

Figure 4 Tree-Pattern-Preserving algorithm extended to preserve multiple trees.

We use the strong form of L-diversity where the fraction of the most frequent values in each equivalence class must be less than 1/L [8]. Thus the maximal probability of privacy breach is 1/L.

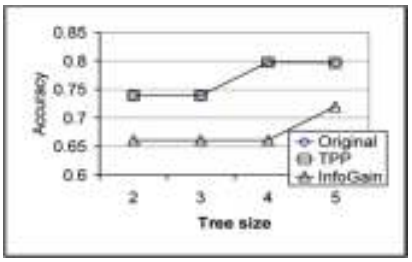


Figure 5 Accuracy of decision trees on Adult data.

Accuracy of tree models. Since both the prediction accuracy and the degree of privacy protection vary with the size of trees, we varied tree size (as the number of leaf nodes) in our experiments. Figure 5 reports the accuracy of decision trees built from sanitized data. The accuracy for trees built from the original data is also reported as the baseline. The results show that the trees built from data sanitized by TPP have higher accuracy than the trees of the same size but built from data sanitized by InfoGain. More importantly, TPP always preserves the decision tree model while InfoGain never preserves the model in all experiments. The accuracy using data sanitized by TPP is the same as that using original data because TPP preserves decision trees. Figure 6 reports the R square of regression trees built from sanitized data. Again, the trees built from TPP have the same mining quality (in terms of R square) as the trees built from the original data. InfoGain does not preserve regression trees and also leads to lower R square.

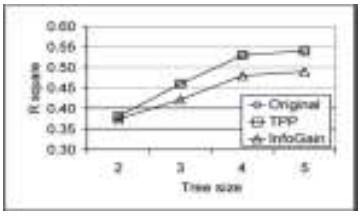


Figure 6 R square of regression trees on Cancer data.

Privacy results. Figures 7 and 8 report K-anonymity results for the two datasets, respectively. K decreases as the tree becomes larger because as the tree grows, more intervals will be generated by TPP and the degree of generalization becomes less. The K values for TPP are slightly worse than those of InfoGain for trees with 4 or 5 leaves, because TPP preserves the tree model and thus does less generalization. This is the price we pay for preserving mining models.

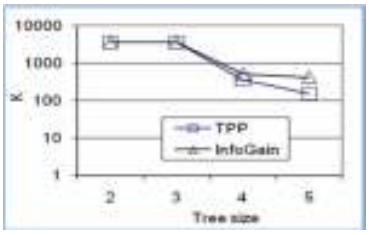


Figure 7 K-anonymity on Adult data.

Volume 8, Issue 6
December 2019

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

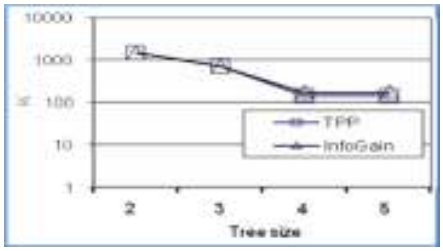


Figure 8 K-anonymity on Cancer data.

5. CONCLUSION

In this paper, mentions a privacy protection technique that preserves decision tree and regression tree models and at the same time protects privacy. We first identify conditions that a privacy protection method must satisfy to preserve the mining models and then design an efficient algorithm that satisfies these conditions.

Experimental results show that our approach not only preserves decision tree and regression tree models, but also leads to better mining quality for several popular mining methods over the sanitized data.

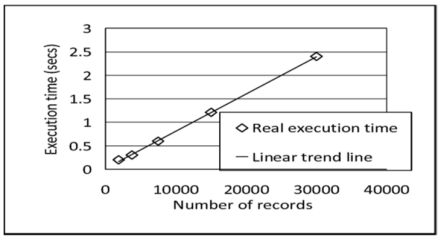


Figure 9 Execution time when varying number of records.

Researchers can use our approach to sanitize their research data and then publish the sanitized data along with mining models. Other researchers can verify the published models using the published data. They can also try other mining methods on sanitized data to solve the same research problem. Application of our approach may potentially reduce both research fraud and encourage sharing of research data.

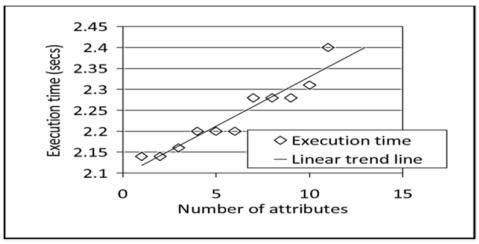


Figure 10 Execution time when varying number of attributes.

As future work, we will investigate whether our approach can be extended to preserve other types of data mining models. Further, it will be interesting to study whether a privacy protection method can preserve the relative order of performance of different mining models. For example, suppose on the original dataset a mining model A (e.g., a decision tree model) is superior to a different mining model B (e.g., a naive Bayesian model), it will be desirable if model A is still better than model B in the sanitized data.

REFERENCES

[1] HAMMANN, F., GUTMANN, H., VOGT, N., HELMA, C., AND DREWE, J. 2010. Prediction of adverse drug reactions using decision tree modeling. Clinic. Pharmacol. Therapeut. To appear.

[2] KUIPERS, T. AND HOEVEN, J. V. D. 2009. Insight into digital preservation of research output in Europe.

[3] LEFEVRE, K., DEWITT, D. J., AND RAMAKRISHNAN, R. 2006b. Workload-Aware anonymization. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 277–286.

[4] LI, N., LI, T., AND VENKATASUBRAMANIAN, S. 2007. t-Closeness: Privacy beyond k-anonymity and l-diversity. In Proceedings of the International Conference on Data Engineering (ICDE'07).

[5] FU, Y.,CHEN, Z.,KORU, A. G., AND GANGOPADHYAY, A. 2009a. A privacy protection technique for publishing data mining models and supporting data. In Proceedings of the Workshop on Information Technologies and Systems Meetings (WITS'09).

[6] FU, Y.,KORU, A. G.,CHEN, Z., AND EMAM, K. E. 2009b. A tree-based approach to preserve privacy of software engineering data and predictive models. In Proceedings of the International Conference on Predictor Models in Software Engineering.

[7] GROSSMAN, R. L., KAMATH, C., KEGELMEYER, P., KUMAR, V., AND NAMBURU, R. (EDS.). 2001. Data Mining for Scientific and Engineering Applications. Kluwer Academic Publishers, Norwell, MA.

[8] XIAO, X. AND TAO, Y. 2006. Anatomy: Simple and effective privacy preservation. In Proceedings of the International Conference on Very Large Databases (VLDB'06). 139–150.

[9] AGGARWAL, C. C. AND YU, P. S. 2004. A condensation approach to privacy preserving data mining. In Proceedings of the International Conference on Extending Database Technology (EDBT'04).

[10] VAIDYA, J., CLIFTON, C., AND ZHU, M. 2005. Privacy Preserving Data Mining. Springer.

[11] LI, X.-B. AND SARKAR, S. 2006a. Privacy protection in data mining: A perturbation approach for categorical data. Inform. Syst. Res. 17, 3, 254–270. LI, X.-B. AND SARKAR, S. 2006b. A tree-based data perturbation approach for privacy-preserving data mining. IEEE Trans. Knowl. Data Engin. 18, 9, 1278–1283.

[12] FU, Y.,CHEN, Z.,KORU, A. G., AND GANGOPADHYAY, A. 2009a. A privacy protection technique for publishing data mining models and supporting data. In Proceedings of the Workshop on Information Technologies and Systems Meetings (WITS'09).

[13] HETTICH, S., BLAKE, C. L., AND MERZ, C. J. 1998. UCI repository of machine learning databases.

[14] MCDAVID, K., SCHYMURA, M. J., ARMSTRONG, L., SANTILLI, L., SCHMIDT, B., BYERS, T., STEELE, C. B.,O'CONNOR, L., SCHLAG, N. C., ROSHALA, W., DARCY, D.,MATANOSKI, G., SHEN, T., AND BOLICK-ALDRICH, S. 2004. Rationale and design of the National Program of Cancer Registries' breast, colon, and prostate patterns of care study. Cancer Causes Control 15, 10, 1057–1066.

[15] NATIONAL PROGRAM OF CANCER REGISTRIES (NPCR). 2010. National center for chronic disease prevention and health promotion.