Use of a Latent Topic Model for Characteristic Extraction from Health Checkup Questionnaire Data*

Y. Hatakeyama1; I. Miyano2; H. Kataoka1; N. Nakajima1; T. Watabe1; N. Yasuda2; Y. Okuhara1

1Center of Medical Information Science, Kochi University Medical School, Kochi, Japan; 2Department of Public Health, Kochi University Medical School, Kochi, Japan

Keywords
Health status, classification, health checkup questionnaire, latent Dirichlet allocation

Summary
Objectives: When patients complete questionnaires during health checkups, many of their responses are subjective, making topic extraction difficult. Therefore, the purpose of this study was to develop a model capable of extracting appropriate topics from subjective data in questionnaires conducted during health checkups.

Methods: We employed a latent topic model to group the lifestyle habits of the study participants and represent their responses to items on health checkup questionnaires as a probability model. For the probability model, we used latent Dirichlet allocation to extract 30 topics from the questionnaires. According to the model parameters, a total of 4381 study participants were then divided into groups based on these topics. Results from laboratory tests, including blood glucose level, triglycerides, and estimated glomerular filtration rate, were compared between each group, and these results were then compared with those obtained by hierarchical clustering.

Results: If a significant (p < 0.05) difference was observed in any of the laboratory measurements between groups, it was considered to indicate a questionnaire response pattern corresponding to the value of the test result. A comparison between the latent topic model and hierarchical clustering grouping revealed that, in the latent topic model method, a small group of participants who reported having subjective signs of urinary disorder were allocated to a single group.

Conclusions: The latent topic model is useful for extracting characteristics from a small number of groups from questionnaires with a large number of items. These results show that, in addition to chief complaints and history of past illness, questionnaire data obtained during medical checkups can serve as useful judgment criteria for assessing the conditions of patients.

1. Introduction

An increase in the number of lifestyle-related diseases in Japan has recently been observed; therefore, more useful methods are needed for analyzing data from health checkups before disease onset. Studies have been conducted to examine the effects of smoking and alcohol consumption using general data from questionnaires conducted during health checkups [1–3]. Starting in 2008, medical checkups in Japan have included an interview section with questions concerning not only alcohol consumption and smoking, but also exercise habits, thereby resulting in an increase in the amount of information obtained from such checkups. In addition to objective data, such as test results, utilization of subjective data from questionnaires may enable more precise evaluation of the patient’s health status and evidence-based health guidance for a variety of lifestyles.

Several studies have been conducted on the evaluation of subjective patient data, such as through the classification of chief complaints [4, 5]. In these studies, descriptions of information regarding patients’ chief complaints or present illnesses were described by the patients themselves, all of whom had subjective symptoms or abnormal test results. As there were many more statements than healthy patients, and each data set always contained a statement showing some abnormality, the classification accuracy was reported to be very high.

Other studies have examined data from health status questionnaires by converting responses into evaluation scores for analysis [6–13]. In these studies, analysis was performed to investigate for the presence of

* Supplementary online material published on our website
http://dx.doi.org/10.3414/ME15-01-0023

© Schattauer 2015
a correlation between evaluation and objective assessment scores and to classify data based on evaluation scores for each type of disease. This method of converting subjective questionnaire data into quantitative data is a useful analytical technique in cases with few items (less than 100) where each item is clearly defined.

Recently, the number of questionnaire items in comprehensive surveys of lifestyle habits has been increasing in Japan. The complexity of relationships between the data from questionnaires makes it difficult to validate the aforementioned conversion of these scores, and prevents classification and evaluation using quantitative questionnaire response patterns. As a result, the large amount of data obtained from these questionnaires becomes a burden on the doctor. Moreover, analysis methods must be validated to increase not only the number of questionnaires, but also the number of questionnaire items. That is, methods appropriate for analyzing a large number of items are necessary before evidence-based health guidance can be implemented. Several studies concerning data mining processes for high-dimensional medical data have recently been conducted [14, 15]. We feel that it is necessary to consider analysis methods for binary data such as questionnaire items.

2. Objectives

The aim of this study was to accurately group patients based on responses to health checkup questionnaires that contained a large number of questions.

Hierarchical classification is one of several approaches used for classifying converted scores; however, this can be difficult when dealing with a large number of questionnaire items. In the classification approach, it is difficult to evaluate how each item may affect the output. When analyzing selection patterns, groups of subjects may have identical response patterns to the questions, making it impossible to create groups of similar subjects. Thus, it is necessary to construct a classification model based on a probability distribution of patient groups and questionnaire items. A probability model called the latent topic model [16, 17] has been studied in the data mining community and applied to both natural language processing (NLP) and image processing. The probability model was originally proposed as generative model for documents in NLP. This model, also referred to as a “bag-of-words,” does not take the order of the words in each document into account, thus assuming an arbitrary frequency distribution, and supposes that the documents include several “topics” which define the frequency distribution of each word. That is, the frequency distribution of the words can be considered a feature of the documents in the model. This model can be applied to image processing [18] by modifying the relationship between the documents and the words.

The aim of this study was to investigate the effectiveness of the latent topic model in identifying characteristics of questionnaire items and their corresponding associations with patients. Therefore, we extracted probability distributions of questionnaire items and topics by modifying documents and words to questionnaire results and items, respectively. The generated topics were then used as features for classification, and the distributions of the items were considered characteristics for evaluating each topic. We obtained each model parameter using latent Dirichlet allocation (LDA) [16, 17]. In the present study, patient groups were defined by topics with the highest probability of occurrence. To assess the appropriateness of this grouping method from the perspective of evidence-based health guidance, we examined the relationship between laboratory measurements obtained during a health checkup, including blood glucose level (GLU), triglycerides (TG), the ratio of one-second forced expiratory volume to forced vital capacity (FEV1%) and estimated glomerular filtration rate (eGFR). In addition, we compared patient groups to test for differences in group allocation based on characteristics identified by the latent topic model.

3. Methods

3.1 Overview

We used a latent topic extraction method to identify topics from questionnaire data from actual health checkups. We assessed the validity of the results by comparing them with results from laboratory tests at the health checkups. To evaluate the effectiveness of this method, we compared the resulting groups of the questionnaire data obtained by hierarchical clustering with those obtained by the latent topic method. The protocol of this study was approved by the Ethics Committee of Kochi University, and all data were anonymized before analysis.

3.2 Data

We used data from health checkups conducted between 2008 and 2010 at the Kochi Medical Checkup Center. Health checkup data were composed of laboratory test results and questionnaire responses from 70,463 records from 40,710 patients. The patients had received checkups once a year, and because some received more than one annual checkup during the study period, only the data from the first checkup during the study period were used for analysis.

All of the following data obtained from the patients were included in the analysis: age; sex; body mass index (BMI); blood pressure; aspartate aminotransferase; alanine aminotransferase; gamma-glutamyl transpeptidase (γ-GT); total protein; albumin; total cholesterol; high-density lipoprotein cholesterol; low-density lipoprotein cholesterol; TG; GLO; hemoglobin A1c; urinalysis (UA); creatinine (Cr); and FEV1%. The data was not corrected for age and sex before performing analysis. In addition, eGFR was calculated using age, sex, and Cr, and used as data for analysis.

Items on the health checkup questionnaire other than those concerning drinking and smoking habits and medical history were in the form of closed-ended questions, with possible responses of “yes” and “no.” The item regarding alcohol consumption had three possible responses (“every day”, “occasionally”, and “never”), as did the item regarding smoking habits (“current smoker”, “past smoker”, and “non-smoker”). The item regarding medical history was an open-ended question. For analysis, these three items were treated as closed-ended questions with responses ex-
pressed, for example, in the form of “every day” or “yes” for alcohol consumption. In total, there were 156 items on lifestyle habits, 33 subjective symptom items, 36 medical family history items, and 45 medical history items. The items used for analysis were those with at least five “yes” responses. Only “yes” responses to health checkup questions were noted.

Because some patients may choose to ignore certain questions, the objects for analysis were extracted under two conditions, that the participants had responded to questions on drinking and smoking habits, and that they had responded “yes” to at least five questions.

Data that did not meet the conditions for the laboratory tests or the health checkup questionnaire were excluded from analysis. Data from a total of 4381 patients (1832 men; mean age [± standard deviation (SD)] 53.2 ± 10.3 years; GLU, 99.7 ± 15.3; TG, 103.7 ± 79.8; FEV1%, 80.1 ± 6.6; eGFR, 73.2 ± 13.7) were used in analysis.

3.3 Extraction of Latent Topics from Health Checkup Data

When analyzing data from health checkup questionnaires, we assume that patients with similar lifestyle habits will have similar responses. Questions selected by a single patient are thought to depend on their lifestyle habits and medical history. This means that some questions selected together will be semantically related. That is, a similar response pattern is defined by a “topic.” This parameter is estimated using data from all subjects. Only questions with “yes” responses were included in the data for analysis; therefore, the data for each subject comprised a collection of questions with “yes” responses. Using LDA [14, 15] with only the “yes” response data enabled extraction of the response patterns of each subject. LDA is a method for calculating multinomial probability distributions used in the latent topic model. A graphical model of this method is shown in Figure 1.

The parameters of a latent topic model are estimated based on the results of responses to all questions. Figure 1 shows a two-stage graphical model generated from data from $M$ input subjects with $N$ items for $k$ topics that were defined in advance. Topic $Z \in \{0, 1, 2, \ldots, k - 1\}$ in target patient $P$ is determined based on the probability $\theta_P(Z)$ as shown in the larger dashed box in Figure 1. Then, the questionnaire item $Q$ is selected by the probability $\phi_P(Z)$ for the selected topic $Z$ as shown in the smaller dashed box in Figure 1. The probabilities $\theta_P$ and $\phi_P$ have a multinomial distribution, while prior distribution is Dirichlet distribution. The $\alpha$ and $\beta$ in Figure 1 represent hyperparameters of the prior Dirichlet distribution. The parameters in this study were defined as 0.5.

The number of topics $k$ must be designated in advance. For the present study, we experimentally designated ($k =$) 30 topics for analysis. Model parameters $\theta_P$ and $\phi_P$ were estimated using the collapsed Gibbs sampler [15] based on the questionnaire response data. This process was implemented in the C# programming language platform.

It is possible to extract characteristics of questionnaire responses for each subject by model estimation. Classification of the subjects is executed using the model parameters. That is, similar groups of patients receiving health checkups were defined as $\theta$ based on calculations using LDA. $\theta_P$ for patient $P$ was denoted by the following equation:

$$\theta_P = \{\theta_P(0), \ldots, \theta_P(Z), \ldots, \theta_P(k - 1)\}$$

$$\text{max_topic}_P = \arg \max \theta_P(Z) \ldots$$

where the probability that topic number $i$ will be selected for patient $P$ is $\theta_P(Z)$.

There are a total of $k$ topics. For patient $P$, the topic number with the highest $\theta_P(Z)$ is $\text{max_topic}_P$. Patients with the same topic number are defined as a group of similar patients. The group of patients corresponding to this topic number $i$ is defined as:

$$C_{\text{total}}(Z) = \{p|\text{max_topic}_P = Z\}, Z = \{0, 1, \ldots, k - 1\} \ldots$$

where $C_{\text{total}}(Z)$ is a set of patients with the same topic number and highest $\theta_P(Z)$.

This definition allowed question characteristics of patient $P$ to be analyzed with $\phi_{\text{max_topic}_P}$ for $\text{max_topic}_P$. This, in turn, enabled analysis of the relationship between questions with a high $\phi_x$ for topic

![Figure 1](image-url)  
**Figure 1** Graphical model representation of latent Dirichlet allocation (the number of questionnaire data: $M$; the number of questionnaire items: $N$; topic: $Z$; questionnaire: $Q$; probability of selecting topic: $\theta$; probability of selecting item: $\phi$; hyperparameters of $\theta$ and $\phi$: $\alpha$ and $\beta$)
Table 1 Number of subjects, percentage of men, mean ± SD age, BMI, and θ for each patient group $C_{\theta(Z)}$

<table>
<thead>
<tr>
<th>Topic no.</th>
<th>n</th>
<th>Male (%)</th>
<th>Age (years)</th>
<th>BMI (kg/m²)</th>
<th>θ (Z)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>67</td>
<td>95.5</td>
<td>47.3 ± 8.5</td>
<td>24.4 ± 3.1</td>
<td>0.26 ± 0.031</td>
</tr>
<tr>
<td>3</td>
<td>76</td>
<td>97.4</td>
<td>51.8 ± 9.3</td>
<td>24.4 ± 3.2</td>
<td>0.26 ± 0.031</td>
</tr>
<tr>
<td>9</td>
<td>78</td>
<td>69.2</td>
<td>53.0 ± 11.1</td>
<td>23.5 ± 4.4</td>
<td>0.26 ± 0.036</td>
</tr>
<tr>
<td>11</td>
<td>88</td>
<td>89.8</td>
<td>60.9 ± 10.0</td>
<td>25.9 ± 4.8</td>
<td>0.25 ± 0.025</td>
</tr>
<tr>
<td>24</td>
<td>295</td>
<td>85.8</td>
<td>57.2 ± 9.8</td>
<td>23.7 ± 2.7</td>
<td>0.27 ± 0.042</td>
</tr>
<tr>
<td>25</td>
<td>571</td>
<td>99.8</td>
<td>61.5 ± 8.6</td>
<td>23.6 ± 2.9</td>
<td>0.27 ± 0.039</td>
</tr>
</tbody>
</table>

Z and laboratory test data for patient group $C_{\theta(Z)}$.

Laboratory test results were compared between each group of subjects using Mann-Whitney U-tests for multiple comparisons, with the p-value adjusted using the Holm method. Actual analytical processing was performed with R version 2.13.0 (R Foundation for Statistical Computing, Vienna, Austria). The level of significance was set at $p = 0.05$.

3.4 Comparison with Hierarchical Clustering

For comparison, hierarchical clustering based on the Manhattan distance, with questions having a response as 1 and not having a response as 0, was then used to group the subjects by topic. Clusters were generated, with the number of clusters $k (= 30)$, the same as that for the number of topics. Actual clustering was performed with R (R Foundation for Statistical Computing).

4. Results

Table 1 shows the number of subjects, number of men, mean age and mean BMI for each patient group $C_{\theta(Z)}$ for topics 1, 3, 9, 11, 24, and 25 (data for all topics are shown in the Appendix). The subjects were divided into 30 groups based on topic (topics 0 to 29). Only the subjects divided into topics 0 to 28 were used in the analyses because topic 29 was only associated with three subjects; these three subjects were thereby excluded.

Table 2 shows the top 5 questions for each topic based on a probability distribution $\phi$ of being selected in topics 1, 3, 9, 11, 24, and 25 (data for all topics are shown in the Appendix).

Table 3 gives the means ± SD for GLU, γ-GT, TG, UA, eGFR, and FEV1% in patient group $C_{\theta(Z)}$. The mean GLU...
score was highest for topic 11 subjects and differed significantly from all the other groups. The mean γ-GT score was highest for topic 3 subjects and differed significantly from topic 2, 5, 6, 12, 13, 14, 15, 16, 18, 19, 20, 21, 22 and 27 subjects. The mean TG score was highest for topic 1 subjects and differed significantly from topic 2, 5, 6, 12, 13, 14, 15, 18, 19, 20, 21, 22, 25 and 28 subjects. The mean UA score was highest for topic 1 subjects and differed significantly from topic 0, 2, 5, 6, 12, 13, 14, 18, 19, 20, 21, 22 and 27 subjects. The mean eGFR score was lowest for topic 11 subjects and differed significantly from topic 0, 2, 5, 6, 7, 12, 13, 14, 15, 18, 19, 20, 21, 22 and 28 subjects. The mean FEV1% score was lowest for topic 9 subjects and differed significantly from topic 14 subjects.

Table 4 shows the classifications obtained with hierarchical clustering of patient data determined to be \( C_{\text{theta}}(Z), Z \in \{9, 11, 25\} \). The classification number refers to the hierarchical classification number. The ratio is the percentage of patients in each hierarchical classification for each topic. Therefore, the sum of the ratios for each topic number is 100%.

### 5. Discussion

Dividing subjects into and evaluating the topic information of each group meant assigning health checkup items that were likely to be selected in each topic with a high \( \phi \) value. For comparison with the questionnaire responses, we selected GLU, γ-GT, TG and eGFR as indexes of glucose metabolism, liver function, lipid metabolism and cardiovascular function, respectively. In addition, eGFR and UA were selected as indexes of renal function.

**Table 3** shows the relationship between questionnaire responses and FEV1%, GLU, and eGFR for topics 9, 11, and 25, respectively. For topics 1, 3, 7, 8, 10, 17, and 26, an association was observed between questionnaire responses concerning dietary habits and γ-GT and TG levels. For each group of subjects, this extraction process was only conducted on responses to items on the health checkup questionnaire; laboratory test results were not considered. If the characteristics of the groups were similar, despite this condition, they were considered to accurately reflect the patient group.

The questionnaire items concerning symptoms associated with loss of cardiovascular function were likely to be selected in topic 9. The patient group \( C_{\text{theta}}(9) \) showed the lowest FEV1% value, but no significant difference was found. Based on these results, topic 9 indicated cardiovascular disorders. In an analogous fashion, topic 11 indicated a history of treatment for lifestyle-related diseases, in consideration of GLU and eGFR. Moreover, topic 25 indicated loss of renal function based on signs of urinary disorder and a lower eGFR.

For topic 24, questions related to hemorrhoids were more likely to be selected. Although not in the top 5 for probability \( \phi \), questions on constipation or past rectal illnesses were often in the top 10.

**Table 4** Results of hierarchical clustering in \( C_{\text{theta}}(Z), Z \in \{9, 11, 25\} \) Classification

- **Table 2** and **Table 3** show the relationship between questionnaire responses and FEV1%, GLU, and eGFR in patients \( C_{\text{theta}}(Z) \) (mean ± SD).
defined in advance. However, reducing the number of topics from 30 to 20 would have prevented extraction of topics such as 9, 24, and 25 in Table 2. This suggests that the total number of topics for the data was appropriate.

Similar items on worsening or improvement of lifestyle habits, subjective symptoms, medical history and other aspects were selected together within topics. This shows that subjects gave responses to similar types of questions. It also shows that the data obtained from our analysis demonstrates that a patient's condition can be projected solely from their responses to questions at a health checkup. The results of our analyses also suggest that questionnaire responses show the same condition as test result data because each subject responded correctly.

We extracted characteristics and divided subjects into groups based on a latent topic model. There are several existing methods for grouping subjects. The clustering method, which aims to generate clusters using some type of measure, has been used in several analytical studies [19–22]. However, whereas LDA can extract the probability of a feature from an input dataset, this type of extraction is difficult using the clustering method. As shown in Table 4, using the clustering method, the subjects in the same $C_{\theta}$ were determined as belonging to several specific groups as opposed to a uniform group. The reason for this is that probability in LDA is related to the frequency distribution of the items used by the clustering method with the input item defined as 0 or 1.

Although 63.6% of the subjects in $C_{\theta}(9)$ were included in Class 1 of the clustering results, the subject groups defined by both approaches differed. One of the reasons for this is that both approaches use different scales: Manhattan distance for the clustering method and occurrence probability for LDA.

The group divided by topic 25 was subdivided into two groups in the clustering method. It is difficult to extract characteristics about renal function directly from the clustering results. The number of questionnaire items concerning renal function is smaller than that of other lifestyle items. Therefore, the clustering results were easily affected by the questionnaire responses regarding other lifestyle items. This explains the discord between the grouping results. The same is true regarding other measurements, that is, the probability of being selected is not uniform for each question. Evaluation using the latent topic model is useful when questions with a small probability of being selected need to be considered. In addition, as the value of $\phi$ for topics can be used to assess the characteristics of each group directly, the latent topic model is more appropriate for evaluating the characteristics of binary data such as health checkup questions.

Factor modeling [23–25] can be used to predict what factors will affect observational data. In this model, it is assumed that observational data is generated by the synthetic quantity of multiple factors, and thus it is used to estimate factor groups and effects. In a factor model, each questionnaire response is assumed to have been generated by a number of factors. In contrast, in the LDA method, as shown in Figure 1, the assumption is that single topics are stochastically allocated to observational data (responses to health checkup questions), and that observational data is generated based on those topics. In other words, it is assumed that the responses for a single subject are generated based on multiple topics, and that each response is generated based on a single topic. The analytical results of the latent topic model were thus easier to evaluate than the results of factor modeling for the input data used in the present study, which included a large number of items.

Table 3 shows that some patient groups had near-abnormal levels for each laboratory test. Therefore, basing patient groups on $C_{\theta}(Z)$ may indicate the possibility of a worsening clinical condition, even if laboratory tests are controlled to normal levels. That is, the extracted characteristics shown in Table 2 suggest clinical conditions that should be kept in mind in evidence-based medical practice.

This study did have some limitations. Some bias regarding the input data may have been present, as one input condition as a criterion for evaluating whether or not a subject had responded appropriately was that they must have responded to the items on alcohol consumption and smoking. As most of the men responded that they drank, questions related to alcohol consumption had a high $\phi$ value for all topics, and were therefore not evaluated. In addition, as the analytical results shown depended only on questions to which responses were given, it was not possible to analyze the effects on test results of questions to which subjects deliberately chose not to respond. Nevertheless, it was still possible to group subjects based solely on the characteristics of reported results, and to assess the relationship between sequential questions and test results.

These results demonstrated that, even for health checkup questionnaires that contain a large number of items, grouping the data based on the latent topic model is possible if the appropriate parameters are set. Moreover, more detailed descriptions of a patient's status can be projected from questionnaires with numerous items. Therefore, characteristics of questionnaire responses evaluated by the latent topic model suggest that evidence-based health guidance based on health checkup data can be provided.

6. Conclusions

The latent topic model appears useful for extracting similar subject responses and features of these subjects from questionnaires with a large number of items. Based on comparisons with laboratory test data, appropriate topics can be generated through the analysis of questionnaires conducted on patients during health checkups.

Acknowledgments

This study was supported by a Grant-in-Aid for Young Scientists (B) (25870488).

References
