Identification of factors associated with hypothyroidism due to radiotherapy in patients with nasopharyngeal cancer
(Case study of nasopharyngeal cancer in one of the hospitals in Jakarta)

D A Lestari¹, S Abdullah¹ and A Rachman²

¹Department of Mathematics, Faculty of Mathematics and Natural Sciences (FMIPA), Universitas Indonesia, Depok 16424, Indonesia
²Division of Hematology and Medical Oncology, Department of Internal Medicine, Faculty of Medicine, Universitas Indonesia, Jakarta 10430, Indonesia

Corresponding author’s email: sarini@sci.ui.ac.id

Abstract. Nasopharyngeal cancer is an abnormal cell growth that develops around the nasopharynx. Treatment of nasopharyngeal cancer patients includes chemotherapy or radiotherapy. Both treatments have side effects in patients. In this study, we will focus on hypothyroidism as a side effect of radiotherapy in the treatment of patients with nasopharyngeal cancer. Hypothyroidism is a condition when the thyroid gland is unable to produce enough thyroid hormone. The main goal of this study is to identify the factors associated with hypothyroidism. To achieve this goal, classification tree and logistic regression methods will be used. Classification tree is used to obtain important variables in the classification of subject classes. Then, logistic regression is used to quantify the risk of variables that appear in the classification tree, hypothyroidism risk factors, and hypothyroidism marker factors. Based on the analysis, it was found that the factors associated in this study were variable symptom, physical sign, smoking habits, gender, age, BMI (Body Mass Index), TSH (Thyroid Stimulating Hormone) and fT4 (free thyroxine) hormones, and also all items on Zulewski score, except items delayed ankle reflex and slow movements. These factors associated tended to increase the risk of hypothyroidism, except for the fT4 hormone and BMI.

Keywords: Classification tree, hypothyroidism, logistic regression, marker factors, risk factors

1. Introduction
Cancer is one of the diseases that cause death with the largest number in Indonesia, and even becomes the second leading cause of death in the world [1]. One type of cancer that has not been widely studied in Indonesia is nasopharyngeal cancer. Nasopharyngeal cancer is an abnormal cell growth that develops around the nasopharynx. In Indonesia, nasopharyngeal cancer ranks fourth as the cancer with the highest incidence after breast cancer, cervical cancer, and lung cancer [2].

In general, the treatment of nasopharyngeal cancer patients is chemotherapy or radiotherapy. Both treatments may have side effects in patients. In this research, we focus on discussing hypothyroidism, one of the side effects of radiotherapy in patients with nasopharyngeal cancer [3]. Hypothyroidism is a condition when the thyroid gland is unable to produce enough thyroid hormone.
Symptoms and signs of hypothyroidism include fatigue, depression, impairment of hearing, constipation, and dry skin. Furthermore, indicators of hypothyroidism also can be known from tests on levels of thyroid stimulating hormone (TSH) and thyroid hormone (fT4). In the very early or mild stage of hypothyroidism, the TSH hormone will usually be high, but the fT4 hormone remains normal [4]. Because the signs and symptoms of hypothyroidism are not specific, it will be useful to have a symptom rating scale to assess the clinical status and potential effects of treatment. Zulewski et al. [5] conducted an evaluation of the results of the Billewicz taking into modern thyroid function tests and producing a Zulewski clinical scores. Evaluation of symptoms and signs with this clinical score serves as an alternative that strengthens the diagnosis of hypothyroidism and is very useful for individual assessment of thyroid function and monitoring in treatment.

Untreated hypothyroidism will affect several health problems, such as heart problems, mental problems, peripheral neuropathy, goiter, nerve and brain disorders, infertility, birth defects, and even lead to decreased quality of life [6]. So, it is important to know and understand the symptoms and signs of hypothyroidism in post radiotherapy patients as early as possible.

Based on this background, we are interested in identifying the factors associated with hypothyroidism along with the characteristics of each factor. We will identify factors from Zulewski score associated with hypothyroidism. We will also identify other hypothyroidism marker factors, such as lab results from the TSH and fT4 hormones, and hypothyroidism risk factors in patients with nasopharyngeal cancer in a hospital in Jakarta.

2. Materials and method

Data on 97 patients with nasopharyngeal cancer who had undergone radiotherapy were obtained from medical records of nasopharyngeal cancer patients in a hospital in Jakarta. Seventy-eight measurements consisting of patient demographic, medical record, physical examination, lab result and Zulewski score were recorded. The use of nonparametric decision tree method allows a large number of measurements, to explain the target variable hypothyroidism status and no distributional assumption is required. Hypothyroidism means that patients with nasopharyngeal cancer who have hypothyroidism, whereas normal means patients with nasopharyngeal cancer who do not have hypothyroidism or are normal.

The methods that will be used in this study are classification tree and logistic regression. The first objective is to identify the factors associated with hypothyroid disease and determine the set of medical information that plays an important role in determining hypothyroid disease. To achieve these objectives the decision tree method will be used because the results of the analysis are easy to interpret and robust to missing values and noise. After obtaining the factors associated with hypothyroid disease, the next step is to determine the characteristics or trends of the factors associated with quantifying the risk of these factors using logistic regression. The software used in this study is R version 1.2.1335 [7].

2.1. Decision tree

A decision tree is a flowchart whose structure is tree like, with each internal node (nonleaf node) showing tests on variables, each branch representing the test results, and each terminal node (leaf node) representing a specific class. The topmost node in the tree is the root node [8]. In this study, the target variable used in the classification model is a categorical variable so that the type of decision tree that will be used is the classification tree with a binary split and two classification classes.

Gini index is a distance criterion based on impurity measurements that measures the divergence between the probability distribution of the target variable value [9]. Suppose C_i is the i class, i = 1, 2 so that C_1 is the event class A and C_2 is the event class B, then this index that measures impurity in the D training data set is denoted as

\[ Gini(D) = 1 - \sum_{i=1}^{m} p_i^2 \] (1)
where $p_i$ is the proportion of data $D$ entered into the $C_i$ class and is estimated by $|C_i|/|D|$ and $m$ is the number of classification classes. In separation with two branches (binary split), we will calculate the number of weights from each partition produced. For example, if binary separation occurs by variable $A$ to $D_1$ and $D_2$, the Gini index value of $D$ data resulting from binary separation by variable $A$ is

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \quad (2)$$

The reduction of impurity that will be generated by binary separation on variable $A$ that is discrete or continuous is

$$\Delta Gini(A) = Gini(D) - Gini_A(D) \quad (3)$$

The variable that maximizes the reduction of impurity or equivalent by having a minimum Gini index is the variable chosen as the separating variable [8].

In this study, the stopping criterion that will be used is the minimum number of observations for the terminal node is 20/3 and the separation criteria threshold is 0.01. This stopping criterion was chosen because it is one of the criteria that forms an optimal tree [10]. The final step in forming a tree is to label the class at each terminal node. Class labeling is done by looking at the results on the terminal node that has the most class frequencies.

The tool used to evaluate the performance of classification methods in recognizing observations and observations from different classes is the confusion matrix. In cases where there are two classes in the datasets that are used, some terminologies could be assigned to differentiate which category is the main interest of the study. ‘Positives’ (P) indicates the total number of individuals who belong to the positive class in the actual data, meanwhile ‘Negatives’ (N) are the remaining ones who belong to the other class. Other terms to be familiar with are ‘True Positives’ (TP) and ‘True Negatives’ (TN), which, respectively, indicate the total number of individuals in the positive and negative class that are correctly labelled by the classifier; ‘False Positives’ (FP), which is the total number of individuals who is predicted to be positive when the actual condition is negative class; and ‘False Negatives’ (FN), which indicates the total number of individuals who is predicted to be negative when the actual condition is positive class.

A popular which is often used is called ‘accuracy’. Accuracy is a measure of model evaluation that shows how much percentage of data classified into the correct class by the classifier.

$$\text{Accuracy} = \frac{TP + TN}{P + N} \quad (4)$$

Accuracy is not a sufficient measure for evaluating a model with an imbalanced data distribution of the class. Therefore, in these cases, the sensitivity and specificity measures can be used as an alternative to the accuracy measures [8].

Sensitivity is a measure of model evaluation that shows the proportion of positive classified data that is identified as being correctly entered into a positive class.

$$\text{Sensitivity} = \frac{TP}{P} \quad (5)$$

Specificity is a measure of model evaluation that shows the proportion of data that is classified as being negatively identified as being correctly classified as negative.

$$\text{Specificity} = \frac{TN}{N} \quad (6)$$

2.2. Logistic regression

Logistic regression is a statistical method to see the effect or tendency of explanatory variables on target variable, where the target variable is categorical variable. The binary logistic regression model is as follows.
where 

\[ E(y) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)} \]



Let \( E(y) \) be the probability when event A occurs, \( \pi \), so \( E(y) = \pi \). The transformation will be carried out so produce the link function as follows.

\[ \text{logit}[\pi(x)] = \ln \left( \frac{\pi(x)}{1 - \pi(x)} \right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \]

In estimating parameters in the logistic regression model, the method commonly used is the maximum likelihood estimator (MLE) method. Interpretation of parameters from the binary logistic regression model using odds ratio.

3. Results and discussion

3.1. Descriptive statistics

Total data in this study were 97 patients with nasopharyngeal cancer who had undergone radiotherapy. Based on figure 1, it could be seen that the number of patients in the hypothyroidism class more than patients in the normal class. There were 63 patients in the hypothyroidism class or around 65 %, while the remaining 34 patients or about 35 % in normal class. The imbalance in the number of patients who are hypothyroid and those who are normal is around 2:1. This shows the imbalance of the patient is not large enough, so it does not affect the results of the analysis.

The hypothyroidism marker factors examined in this study were Zulewski score [5], TSH (Thyroid Stimulating Hormone) and fT4 (free thyroxine) hormone [4]. Furthermore, the hypothyroidism risk factors examined in this study were smoking, gender and age [11], also BMI (Body Mass Index) [12].

Based on the table 1, with a significance level of 0.05, it was found that the Zulewski variables that had a relationship with the subject class, that is, having a p-value less than 0.05 were the diminished sweating, hoarseness, paraesthesia, dry skin, constipation, weight increase, and coarse skin. Therefore, Zulewski variables will be further analysed to determine the characteristics and quantify the risk of each variable for hypothyroidism using logistic regression. Delayed ankle reflex variable was not included in the independence test because of the uniformity of the data on the variable.

![Figure 1. Percentage of the number of subjects per class](image)
Table 1. Statistic descriptive for Zulewski score.

| Variable                | Subject class | P-value independence test |
|-------------------------|---------------|----------------------------|
|                         | Hypothyroidism | Normal |                         |
|                         | n (%)         | n (%)  |                           |
| Diminished sweating     | Yes           | 41 (97.62) | 1 (2.38) | 0.00 |
|                         | No            | 22 (40.00) | 33 (60.00) |       |
| Hoarseness              | Yes           | 43 (95.56) | 2 (4.44) | 0.00 |
|                         | No            | 20 (38.46) | 32 (61.54) |       |
| Symptom                 | Parasthesia   | Yes     | 40 (97.56) | 1 (2.44) | 0.00 |
|                         | No            | 23 (41.07) | 33 (58.93) |       |
| Symptom                 | Dry skin      | Yes     | 48 (96.00) | 2 (4.00) | 0.00 |
|                         | No            | 15 (31.91) | 32 (68.09) |       |
| Symptom                 | Constipation  | Yes     | 16 (94.12) | 1 (5.88) | 0.01 |
|                         | No            | 47 (58.75) | 33 (41.25) |       |
| Symptom                 | Impairment of hearing | Yes | 39 (67.24) | 19 (32.76) | 0.56 |
|                         | No            | 24 (61.54) | 15 (38.46) |       |
| Symptom                 | Weight increase | Yes | 40 (88.89) | 5 (11.11) | 0.00 |
|                         | No            | 23 (44.23) | 29 (55.77) |       |
| Symptom                 | Slow movements | Yes | 3 (100.00) | 0 (0.00) | 0.55 |
|                         | No            | 60 (63.83) | 34 (36.17) |       |
| Physical                | Delayed ankle reflex | Yes | 0 (0.00) | 0 (0.00) |       |
|                         | No            | 63 (64.95) | 34 (35.05) |       |
| Physical                | Coarse skin   | Yes     | 26 (96.30) | 1 (3.70) | 0.00 |
|                         | No            | 37 (52.86) | 33 (47.14) |       |
| Physical                | Periorbital puffiness | Yes | 10 (90.91) | 1 (9.09) | 0.09 |
|                         | No            | 53 (61.63) | 33 (38.37) |       |
| Physical                | Cold skin     | Yes     | 5 (83.33) | 1 (16.67) | 0.66 |
|                         | No            | 58 (63.74) | 33 (36.26) |       |

Table 2. Statistic descriptive for TSH and fT4 hormones.

| Variable | Subject class | Mean     | Std. deviation | P-value t test |
|----------|---------------|----------|----------------|----------------|
| TSH      | Hypothyroidism| 4.13     | 5.48           | 0.34           |
|          | Normal        | 3.12     | 3.43           |                |
| fT4      | Hypothyroidism| 1.07     | 0.20           | 0.22           |
|          | Normal        | 1.31     | 1.14           |                |

Based on table 2, it could be seen in the data that there are no significant differences in the average of TSH and fT4 hormones in the group of patients who have hypothyroidism and normal. However, based on the mean column in that table, it could be seen that there is a tendency that the average value of TSH hormone in the hypothyroidism class is higher than the average value of TSH hormone in the normal class. In contrast, the average value of fT4 hormone in the hypothyroidism class tends to be smaller than the average value of fT4 hormone in the normal class. In addition, the standard deviation of the TSH hormone variable is large compared to the standard deviation of the fT4 hormone variable. This indicates that the TSH hormone variable varies greatly.
Based on figure 2, it was found that the TSH and fT4 variables have enough outliers so that from the boxplot the average TSH and fT4 variables in the hypothyroidism and normal subject classes are almost the same. Many outliers on TSH and fT4 variables data make the distribution of data abnormal so that the results of the t test are invalid because they do not require the assumptions of the t test, normality. Because the TSH and fT4 variables are variables from the lab results so the authors do not carry out imputation or data cleaning, because the lab results information provided is quite sensitive and based on actual measurements. Therefore, this analysis will proceed by looking at the tendency and risk of the TSH and fT4 variables with using logistic regression.

Based on table 3, it could be seen that with a significance level of 0.05, smoking, gender, and age have no significant relationship with hypothyroidism. This can occur because the number of samples in this study is quite small. However, this is not a problem because the analysis will continue by looking at the tendency and risk of each variable using logistic regression.

The mean column in table 4 shows the difference in mean BMI scores for each subject class. Based on table 4, it could be seen that the average BMI value of subjects who have hypothyroidism is lower compared to normal subjects. With a significance level of 0.05, the BMI variable has a difference in mean values between subject classes because the p-value of the BMI variable is 0.01 less than 0.05 so it could be said the BMI variable affects the target variable (subject class) individually.

![Figure 2. Boxplot of TSH and fT4 hormones variable](image)

**Table 3. Statistic descriptive for risk factors (categorical)**

| Variable | Subject class          | P-value independence test |
|----------|------------------------|---------------------------|
|          | Hypothyroidism | Normal |
| Smoking  | Yes 25 (76 %)          | 8 (24 %) | 0.11 |
|          | No 38 (59 %)          | 26 (41 %) |
| Gender   | Female 23 (66 %)       | 12 (34 %) | 0.91 |
|          | Male 40 (65 %)         | 22 (35 %) |
| Age      | ≥ 60 10 (71 %)         | 4 (29 %) | 0.77 |
|          | < 60 53 (64 %)         | 30 (36 %) |
3.2. Analysis data

Based on table 5, it could be seen that the model or scheme that has the highest accuracy is the Zulewski score model (simplification item). Therefore, this study will discuss in detail the classification tree model of Zulewski score (simplification items). The classification tree method provides results in the form of relationship information about the explanatory variable to the target variable including the characteristics of the explanatory variable in the class classification of the subject.

The rules produced by the classification tree on figure 3 are as follows.

- Someone classified will enter the hypothyroidism class if he has more than three Symptom or has less than three Symptom but has a Physical Sign of more than or equal to one.
- Someone classified will enter the normal class if he has less than three Symptom and does not have Physical Sign.

| Scheme | Information | Training data | Important variable |
|--------|-------------|----------------|--------------------|
| 1      | Medical record when diagnosing nasopharyngeal cancer | 0.90 (0.82; 0.95) | 0.89 | 0.91 | sweating diminished and numb |
| 2      | Physical examination | 0.75 (0.65; 0.83) | 0.71 | 0.82 | BMI and temperature |
| 3      | Laboratorium result | 0.84 (0.75; 0.90) | 0.90 | 0.71 | SGPT, fT4, Ur, dan Cr |
| 4      | Zulewski score (Individual Item) | 0.93 (0.86; 0.97) | 0.92 | 0.94 | dry skin dan constipation |
| 5      | Zulewski score (Simplification Item) | 0.97 (0.91; 0.99) | 1.00 | 0.91 | symptom dan physical sign |
| 6      | All information (without Zulewski score) | 0.90 (0.82; 0.95) | 0.89 | 0.91 | sweating diminished and numb |
| 7      | All information (with Zulewski score individual item) | 0.93 (0.86; 0.97) | 0.92 | 0.94 | dry skin dan constipation |
| 8      | All information (with Zulewski score simplification item) | 0.97 (0.91; 0.99) | 1.00 | 0.91 | symptom dan physical sign |

Table 4. Statistic descriptive for risk factors (numerical).

| Variable | Subject class | Mean | Std. deviation | P-value t test |
|----------|---------------|------|----------------|--------------|
| BMI      | Hypothyroidism| 20.40| 3.03           | 0.01         |
|          | Normal        | 21.94| 2.46           |              |
From model Zulewski score (simplification item) obtained the important variable (or it could be said factors associated with hypothyroidism) are symptom and physical sign. After getting the variables associated with hypothyroidism using classification tree, the next step is to find out the characteristics and quantify the risk of these variables using logistic regression.

Based on table 6, it was found that the odds ratio for the symptom variable is 32.33. This states that the addition of one unit of symptom variable is partially in line with the increased risk of someone developing hypothyroidism by 32.33 times than before. Then, the odds ratio value for the physical sign variable is 12.70. This states that the addition of one unit of physical sign variable is partially in line with the increased risk of someone developing hypothyroidism by 12.70 times than before.

Based on table 7, some information can be obtained as follows.

- The tendency of someone who smokes to developing hypothyroidism is about 2.14 times compared to someone who doesn't smoke.
- The tendency of someone with age ≥ 60 to developing hypothyroidism is about 1.42 times compared to someone with age < 60.
- The tendency of someone who is female to developing hypothyroidism is about 1.05 times compared to someone who is male.
- Addition one unit of BMI tends to reduce the risk of developing hypothyroidism by 0.82 times than before.

![Classification tree](image)

**Figure 3.** Classification tree from scheme Zulewski score simplification item

**Table 6.** Results of logistic regression analysis for important variable in classification tree model.

| Explanatory variable | Coefficient $\beta$ | Odds ratio |
|----------------------|---------------------|------------|
| Symptom              | 3.47                | 32.33      |
| Physical Sign        | 2.54                | 12.70      |
The odds ratio used by the authors in interpreting parameters is the odds ratio with 0.1 units. This is due to the numerical scale TSH and fT4 hormones data characteristics and the distribution of values on the TSH and fT4 hormones data on the tithe scale. Based on table 8, it could be seen that the TSH and fT4 hormones have an odds ratio that is almost equal to one. It states that an increase in one unit of TSH or fT4 hormones is not associated with an increase or decrease in someone’s risk of developing hypothyroidism. However, based on the odds ratio values obtained through logistic regression analysis, the odds ratio for the TSH hormone is 1.01 so it could be said that the increase in 0.1 unit of TSH hormone tends to be in line with the increased risk of someone developing hypothyroidism. Conversely, the odds ratio for the fT4 hormone is 0.92 so it could be said that the increase in 0.1 unit of the fT4 hormone tends to be in line with the decreased risk of someone developing hypothyroidism.

Based on table 9, the odds ratio value for each item Zulewski score (except delayed ankle reflex items) is greater than one. Because the Zulewski score only has two values, namely 1 (Yes) and 0 (No), it could be said that someone who has a health disorder as contained in the Zulewski clinical score tends to be at risk of developing hypothyroidism compared to someone who does not have the health disorder. The variables of Zulewski score associated with hypothyroidism are diminished sweating, hoarseness, paraesthesia, dry skin, constipation, impairment of hearing, weight increase, coarse skin, periorbital puffiness, and cold skin. The delayed ankle reflex variable was not included in the logistic regression analysis because there was uniformity of data on this variable, i.e. all observations chose “No” for the delayed ankle reflex variable. This uniformity of data also occurs in the slow movements variable, i.e. there are only 3 observations that choose “Yes”. That is what causes the odds ratio of the slow movements variable to be very large and unnatural. Thus, it can be said that the variable delayed ankle reflex and slow movements are not associated with hypothyroidism.

3.3. Discussion
Zulewski et al. [5] in their study said that the variables associated with hypothyroidism were all items that built up Zulewski clinical score. This is quite in line with the study of the authors, but in this study the slow movements and delayed ankle reflex variables is not associated with hypothyroidism. This is due to differences in the data and models used, or possible differences in biological factors from the research subjects. Koulouri et al. [4] said that an indication of someone developing hypothyroidism is when TSH hormone levels are high and levels of fT4 hormone are low or normal. This is in line with the results of this study that an increase in TSH hormone increases the risk of developing hypothyroidism and an increase in the hormone fT4 reduces the risk of developing hypothyroidism.

| Table 7. Results of logistic regression analysis for hypothyroidism risk factors. |
| Explanatory variable | Coefficient β | Odds ratio |
|-----------------------|---------------|------------|
| Smoking (Yes/No)      | 0.76          | 2.14       |
| Age (≥ 60/< 60)       | 0.35          | 1.42       |
| Gender (Female/Male)  | 0.05          | 1.05       |
| BMI                   | -0.20         | 0.82       |

| Table 8. Results of logistic regression analysis for hypothyroidism marker factors (TSH and fT4 hormones). |
| Explanatory variable | Coefficient β | Odds ratio |
|----------------------|---------------|------------|
| TSH                  | 0.06          | 1.06       |
| fT4                  | -0.85         | 0.43       |

Table 8. Results of logistic regression analysis for hypothyroidism marker factors (TSH and fT4 hormones).
Table 9. Results of logistic regression analysis for hypothyroidism marker factors (Zulewski score).

| Explanatory variable                  | Coefficient β | Odds ratio |
|---------------------------------------|---------------|------------|
| Diminished sweating                   | 4.12          | 61.50      |
| Hoarseness                            | 3.54          | 34.40      |
| Paraesthesia                          | 4.05          | 57.39      |
| Symptom                               |               |            |
| Dry Skin                              | 3.94          | 51.20      |
| Constipation                          | 2.42          | 11.23      |
| Impairment of hearing                 | 0.25          | 1.28       |
| Weight increase                       | 2.31          | 10.09      |
| Slow movements                        | 16.00         | 8869242.94 |
| Delayed ankle reflex                  | -             | -          |
| Physical                              |               |            |
| Coarse skin                           | 3.14          | 23.19      |
| Periorbital puffiness                 | 1.83          | 6.23       |
| Cold skin                             | 1.05          | 2.84       |

Some risk factors for hypothyroidism are someone over the age of 60 years, female, and smoking [11]. This is in line with the results of this study that the three risk factors mentioned before do tend to increase the risk of hypothyroidism. Diab et al. in their study said the greater the value of BMI has an impact on increasing someone’s risk of developing hypothyroidism [12]. However, the authors found that the increase in the BMI variable is in line with the reduced risk of someone developing hypothyroidism. Based on the descriptive statistics of the BMI variable described previously, observations that have hypothyroidism tend to have a lower BMI compared to observations with normal status. Some things that might explain the different conditions of the results of this study are the characteristics of the patient, the model used, or other possibilities.

4. Conclusion

Factors associated with hypothyroidism consist of important variables resulting from classification tree analysis, hypothyroidism risk factors, and hypothyroidism marker factors. These factors include variable symptom, physical sign, smoking, gender, age, BMI (Body Mass Index), TSH (Thyroid Stimulating Hormone) and fT4 (free thyroxine) hormones, and all Zulewski score items, except items slow movements and delayed ankle reflex. These factors associated tended to increase the risk of hypothyroidism, except for the hormones fT4 and BMI. Someone classified will enter the hypothyroidism class if he has more than three symptoms or has less than three symptoms but has a physical sign of more than or equal to one. Moreover, someone classified will enter the normal class if he has less than three symptoms and does not has a physical sign.

References

[1] World Health Organization 2018 Cancer available at https://www.who.int/newsroom/factsheets/detail/cancer
[2] Adham M et al. 2017 Pedoman Nasional Pelayanan Kedokteran Kanker Nasofaring (Jakarta: Kementrian Kesehatan Republik Indonesia)
[3] Boomsma M J, Bijl H P, Christianen M E, Beetz I, Chouvalova, O, Steenbakkers R J, Wolffenbutel B H, Oosting S F, Schilstra C and Langendijk J A 2012 Int. J. Radiat. Oncol. Biol. Phys. 84 351-356
[4] Koulouri O and Gurnell M 2013 Clin. Med. 13 282-6
[5] Zulewski H, Muller B, Exer P, Miserez A R and Staub J J 1997 *J. Clin. Endocrinol. Metab.* **82** 771-6

[6] Mayo Clinic Staff 2018 *Hypothyroidism (Underactive Thyroid)* available at https://www.mayoclinic.org/diseases-conditions/hypothyroidism/symptoms-causes/syc-20350284

[7] R Core Team 2019 *R: A Language and Environment for Statistical Computing* available at https://www.R-project.org/

[8] Han J and Kamber M 2006 *Data Mining: Concepts and Techniques* 2nd edition (United Kingdom: Elsevier)

[9] Rokach L and Maimon O Z 2015 *Data Mining with Decision Trees: Theory and Applications* (New Jersey: World Scientific)

[10] Therneau T, Atkinson B and Ripley B 2019 Package ‘Rpart’: Recursive Partitioning and Regression Trees available at https://cran.r-project.org/web/packages/rpart/rpart

[11] Data and Information Center of the Ministry of Health of the Republic of Indonesia 2015 *Situasi dan Analisis Penyakit Hipotiroid* (Jakarta: LIPI)

[12] Diab N, Daya N R, Juraschek S P, Martin S S, McEvoy J W, Schultheiß U T, Kottgen A and Selvin E 2019 *Sci. Rep.* **9** 1-8