MIND dataset for diet planning and dietary healthcare with machine learning: Dataset creation using combinatorial optimization and controllable generation with domain experts

Changhun Lee [*] [1], Soohyeok Kim [*] [1], Sehwa Jeong[1], Jayun Kim[2], Yeji Kim[2],
Chiehyeon Lim [†] [1], Minyoung Jung [†] [3]

[1]Ulsan National Institute of Science and Technology (UNIST)
{messy92, sooo, jsh0746, chlim}@unist.ac.kr
[2]Kosin University Gospel Hospital
{jydk6557, kimhana0419}@naver.com
[3]Kosin University College of Medicine
{my.jung}@kosin.ac.kr

Abstract

Diet planning, a basic and regular human activity, is important to all individuals. Children, adults, the healthy, and the infirm all profit from diet planning. Many recent attempts have been made to develop machine learning (ML) applications related to diet planning. However, given the complexity and difficulty of implementing this task, no high-quality diet-level dataset exists at present. Professionals, particularly dietitians and physicians, would benefit greatly from such a dataset and ML application. In this work, we create and publish the Korean Menus–Ingredients–Nutrients–Diets (MIND) dataset for a ML application regarding diet planning and dietary health research. The nature of diet planning entails both explicit (nutrition) and implicit (composition) requirements. Thus, the MIND dataset was created by integrating input from experts who considered implicit data requirements for diet solution with the capabilities of an operations research (OR) model that specifies and applies explicit data requirements for diet solution and a controllable generative machine that automates the high-quality diet generation process. MIND consists of data from 1,500 South Korean daily diets, 3,238 menus, and 3,036 ingredients. MIND considers the daily recommended dietary intake of 14 major nutrients. MIND can be easily downloaded and analyzed using the Python package dietkit accessible via the package installer for Python. MIND is expected to contribute to the use of ML in solving medical, economic, and social problems associated with diet planning. Furthermore, our approach of integrating data from experts with OR and ML models is expected to promote the use of ML in other fields that require the generation of high-quality synthetic professional task data, especially since the use of ML to automate and support professional tasks has become a highly valuable service.

[*]Equal contribution.

[†]Corresponding author.

Submitted to the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks. Do not distribute.

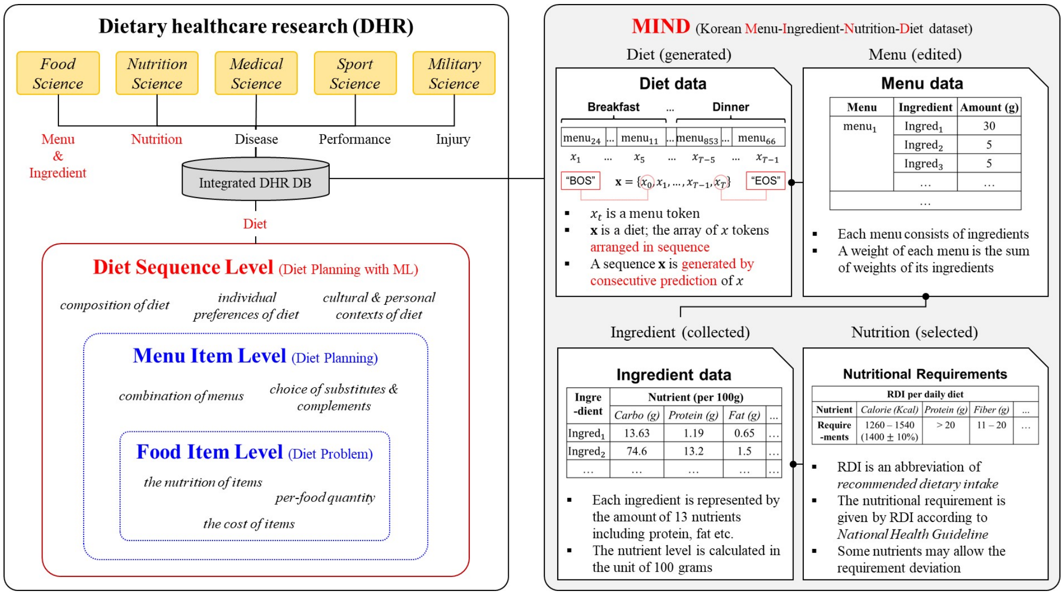
1 Introduction

Diet is “the sum of foods consumed by a person or other organism” [46], and diet planning is a regular human activity. The term “meal” implies consumed foods in general, and the term “diet” is used to indicate the combination of food menus planned for a specific purpose such as nutritional satisfaction, allergen avoidance, or weight control [18, 37]. Given that a diet is necessary for all individuals, diet planning has emerged as a core function of dietary healthcare research (DHR) in diverse disciplines that include food technology [42, 68, 69], nutrition management [15], clinical medicine [77], sports science [6, 33], and military nutrition [50, 25]. A single diet can be defined as a sequence of menus; diet planning involves the consideration of menus, ingredients, and nutrients (see Figure 1). A menu item is the complete product of cooked foods. For example, “a salad” is food and “ricotta cheese salad” is on the menu. Individuals usually consume end-products, not raw foods, and "menu" corresponds to the end product. “Ricotta cheese salad” consists of ingredients such as ricotta cheese, lettuce, and balsamic vinegar; and each ingredient contains several nutrients such as protein, fat, iron, sodium, etc. Therefore, any single diet can be hierarchically expressed with respect to menu-level, ingredient-level, or nutrient-level representations.

Diet planning is an advanced issue of the traditional "diet problem", the problem of optimizing quantities of foods and ingredients. The diet planning problem involves assessment of menus rather than foods. The solution to this problem is the optimization of the quantity of each menu with the simultaneous attainment of the optimal combination of menus (refer to Section 2 and Appendix A.1 for further details on the diet problem and diet planning). Recently in the healthcare field, researchers have attempted to define a health-related diet planning problem and to solve this problem using machine learning (ML). A major interest of medical DHR with ML is the design of a diet that counters disease-related factors [77, 39, 62, 2], and the ML studies of sports and military DHR focus on diets that strengthen physical abilities and metabolic controls [26, 16]. Despite the importance of ML application in academia and practice, studies in ML-based DHR are challenging because of the insufficiency of data. Figure 1 illustrates how DHR studies have been conducted based on the data of diet + X (e.g., menu, ingredient, or nutrition) configurations. Most of these previous studies have evaluated the physiological changes in subjects consuming different foods or have focused on recommending the consumption of specific foods based on perceived benefit. This indicates that diet data are the main source of information in those studies. However, a sufficiently large benchmark diet dataset that is accessible to the public does not yet exist. [17, 24, 57, 79]. This lack of a diet-level dataset may be the reason that most dietary studies have been based on operations research (OR) modeling instead of the ML approach that requires a dataset for training.

Several reasons exist for the lack of a diet-level dataset. From a data perspective, the diet can be defined as a set of menu items or food items arranged in a sequence, e.g., appetizer, main course, and dessert, for a specific purpose (see Figure 1). Obtaining a large quantity of diet data from current consumption practices may appear to be relatively simple. However, actual diet data have significant data quality issues. Our previous study provides evidence of this [35, 28]. While we were able to obtain an actual diet dataset that was created and used by public institutes and professional dietitians in South Korea, difficulty in use of this as a benchmark dataset arose for two reasons. First, the nutritional quality of each diet was inadequate. The first objective of dietary studies is to meet nutritional requirements according to age or other conditions, and necessary guidelines are clearly delineated by nutrition science. Surprisingly, many of the diets provided by public institutes did not meet these requirements. Many dietitians believe that this is an unavoidable reality because of the high complexity and difficulty of diet planning. Designing a diet plan is indeed complex and difficult because of its combinatorial optimization nature, which represents an NP-hard problem [74, 51]. For example, a breakfast plan with a combination of 100 menu items will consist of approximately 10^8 options, supposing that a breakfast contains five menu items.

Second, the available datasets are insufficient in size. Usually, a unit of data in a diet dataset is one daily diet. Therefore, yearly data only contain approximately 300 examples, limiting the composition patterns of the diets. Additionally, diet planning involves substantial knowledge of food and nutrition. Understanding the context, e.g., religious beliefs and cultural


Dietary healthcare research (DHR)
Food Science
Nutrition Science
Medical Science
Sport Science
Military Science
Menu & Ingredient
Nutrition
Disease
Performance
Injury
Integrated DHR DB
Diet
Diet Sequence Level (Diet Planning with ML)
composition of diet
individual preferences of diet
cultural & personal contexts of diet
Menu Item Level (Diet Planning)
combination of menus
choice of substitutes & complements
Food Item Level (Diet Problem)
the nutrition of items
per-food quantity
the cost of items
MIND (Korean Menu-Ingredient-Nutrition-Diet dataset)
Diet (generated)
Menu (edited)
Diet data
Breakfast
...
Dinner
menu_{24} ... menu_{11} ... menu_{853} ... menu_{66}
x_1 ... x_5 ... x_{T-5} ... x_{T-1}
"BOS"
$\mathbf{x} = \{x_0, x_1, \dots, x_{T-1}, x_T\}$
"EOS"
x_t is a menu token
$\mathbf{x}$ is a diet; the array of x tokens arranged in sequence
A sequence $\mathbf{x}$ is generated by consecutive prediction of x
Menu data
Each menu consists of ingredients
A weight of each menu is the sum of weights of its ingredients
Ingredient (collected)
Nutrition (selected)
Ingredient data
Each ingredient is represented by the amount of 13 nutrients including protein, fat etc.
The nutrient level is calculated in the unit of 100 grams
Nutritional Requirements
RDI is an abbreviation of recommended dietary intake
The nutritional requirement is given by RDI according to National Health Guideline
Some nutrients may allow the requirement deviation

Menu	Ingredient	Amount (g)
menu_1	Ingred_1	30
	Ingred_2	5
	Ingred_3	5
	…	…
…		

Ingre-dient	Nutrient (per 100g)			
	Carbo (g)	*Protein (g)*	*Fat (g)*	...
Ingred_1	13.63	1.19	0.65	…
Ingred_2	74.6	13.2	1.5	…
…	…	…	…	…

RDI per daily diet				
Nutrient	*Calorie (Kcal)*	*Protein (g)*	*Fiber (g)*	...
Require-ments	1260 – 1540 (1400 ± 10%)	> 20	11 – 20	…

Figure 1: The scope of our study (left) and structure of the MIND dataset (right). The approaches in the blue boxes are used by most OR studies, which are based on the formulation of explicit requirements of diet planning; the approach in the red box is extended to learn implicit patterns in diets through ML. This figure shows the spectrum from existing works, primarily using an OR approach to confront the diet problem and diet planning to our ML-based approach to address these issues. In summary, all previous studies on diet planning consider ingredient and menu-level information, but diet-level planning should involve the compositional patterns of menus in diets. In addition, existing ML studies on dietary healthcare also consider only the ingredient and menu levels. The proposed MIND dataset is the first dataset that integrates all of the hierarchical relationships between diets-menus, menus-ingredients, and ingredients-nutrition.

orientation, and health and development issues, e.g., growth, aging, and the pathogenesis of chronic diseases, is also of prime importance [45, 47]. This knowledge must be treated as constraints when generating diets, but only some of these topics have an explicit guide for specifying nutritional and other dietary requirements. No guidelines exist for the remaining topics because the guidelines and topics are related to implicit requirements that include the composition of a diet. As a result, professional dietitians employed in government or daycare centers often copy and edit existing diets that are poorly crafted (see Section 4), and this emulation behavior adversely impacts the quality and size of available diet datasets. Similarly, although medical doctors and dietitians in large hospitals should design specialized diet plans for inpatients, few inpatients receive these services. Last, diet planning in the home is usually unsystematic, contributing to the low quality and insufficient size of the available benchmark dataset. Therefore, the focus of our study is data augmentation using synthetic diets of high quality to construct a benchmark dataset for ML-based diet planning applications and DHR.

To generate synthetic diets of high quality, we initially performed the task of diet generation by redefining the traditional OR diet planning problem as an ML one, a controllable generation problem as described in Section 2. Accordingly, we devised an OR–Xperts–ML (ORxML) framework that integrates input from experts with the capabilities of OR and ML modules (see Section 3). Each OR, Expert, and ML module is responsible for the initialization, evaluation, adjustment, and control of diet generation. The specific process involves the formulation of a combinatorial optimization OR model to generate synthetic diets as a means of satisfying explicit nutrient requirements. Next, we recruited experts, professional dietitians, to evaluate and adjust the initial data in terms of implicit requirements. These implicit requirements are criteria that cannot be specified in the combinatorial optimization model. An example of these requirements is the essential dietician task of assessing the composition of a diet based on its implicit and contextual nature. This is critical to make the diet recipients accept and enjoy menus with high nutritional quality. See Appendix for further details on the compositional quality of diets. Without this consideration, feasible solutions for diet planning cannot be provided in practice. Last, we developed a controllable diet generation machine to: (a) ensure composition compliance by learning the data patterns constructed by the OR model and experts, (b) enhance nutrition by approximating an optimal policy to maximize the nutrient rewards, and (c) automatically augment the data by executing an optimal policy and generating synthetic diets.

With the diets generated by the ORxML framework, we created the *Menu–Ingredient–Nutrient–Diet* (MIND) dataset for diet planning and DHR with ML and introduce this dataset in this study. Figure 1 shows the MIND dataset that consists of 1,500 daily diets, 3,238 menus, and 3,036 ingredients. Satisfaction of the nutritional intake requirements for 14 major nutrients was a significant consideration. The original sources of the menu items, ingredients, and nutrient information are the public databases of South Korean government organizations that are responsible for ensuring the country's nutrition standards, and the diet data were created by the authors from the beginning using the ORxML framework. The quality of the diets was validated by dietitians and physicians, and we received approval from the government organizations responsible for determining nutrition quality in South Korea (e.g., the Ministry of Food and Drug Safety and the Rural Development Administration) to distribute the MIND dataset. The MIND dataset can be downloaded and subsequently analyzed easily using the Python package called *dietKit*, which is accessible via the package installer for Python.

This work is original research with academic merit and practical implications as illustrated in Figure 1. Diet planning is an important problem that should be solved with ML but could not be addressed in this way due to the lack of datasets for this data-driven approach. To the best of our knowledge, this work is the first to create and publish a large-scale and high-quality diet-level dataset for diet planning and DHR using ML. Section 2 explains the methodological background more thoroughly. In addition, this work represents a first attempt to develop a framework for generating high-quality synthetic data for professional tasks. Section 3 explains the ORxML framework in detail. In Section 4, we discuss how the quality of the MIND dataset was evaluated via a series of experiments to demonstrate the significance of the three modules, the OR model, the knowledge and experience of experts, and the ML model. The final outcome of the MIND dataset is described in Section 5. Our work has already started to create an impact. In Section 6, we discuss ML applications of our dataset as a means of assisting dietitians, medical doctors, and the public in their diet planning and related healthcare tasks. In Section 7, we discuss how the ORxML framework can be applied to constructing high-quality synthetic data involving professional tasks in other domains.

2 Background and Literature Review

The academic concepts and definitions necessary to understand our research are briefly discussed in this section. Each of the two subsections defines the diet planning problem and its recent paradigm with the support of ML.

Diet planning problem The concept of the *diet problem*, highlighted by Dantzig [14], was motivated by the United States Army's desire to meet the nutritional requirements of military personnel in the field while minimizing the cost of implementing the endeavor [3]. The prototype study of the diet problem was published in 1945 when George Stigler, who later received the Nobel Prize, presented an economical diet model [64]. Stigler regarded the diet problem as a scenario involving continuous optimization to identify optimal quantities of food items; thus, a linear programming approach was adopted. However, Stigler's approach was later criticized as impractical by subsequent economists and operation researchers. Most criticisms centered on the optimization units. Smith [61] and Smith [60] explained that the linear programming solution, i.e., using an optimal set of food items, was "unpalatable" because the linear models exemplified "one-dish meals" similar to animal feed blends rather than those fit for a "daily human diet." Similarly, Peryam [49] and Eckstein [19] also disapproved of the linear programming approaches. Their contention was that a diet is optimized at the level of food items or ingredients rather than at the level of menu items or recipes; the solutions of the linear programming were viewed as raw materials and not as end-products. This view is critical, as humans do not consume a specific quantity of each food unit but rather the end-product as a whole unit. Subsequent to this wave of criticism of Stigler's approach, a new type of diet problem, i.e., diet planning, has emerged. The diet problem in this case has been formulated as a combinatorial optimization problem. Consequently, most researchers have applied mixed-integer programming (MIP) to solve the diet problem. Diet planning research has since focused on formulating the diet problem in the MIP form [59, 36, 22] as follows:

$$\max_{\mathbf{X}} \text{ total nutrition} = \sum_{t=1}^{T} \left(\mathbf{A}^{\mathrm{T}}\mathbf{X}\right)_{nt} \tag{1}$$

subject to

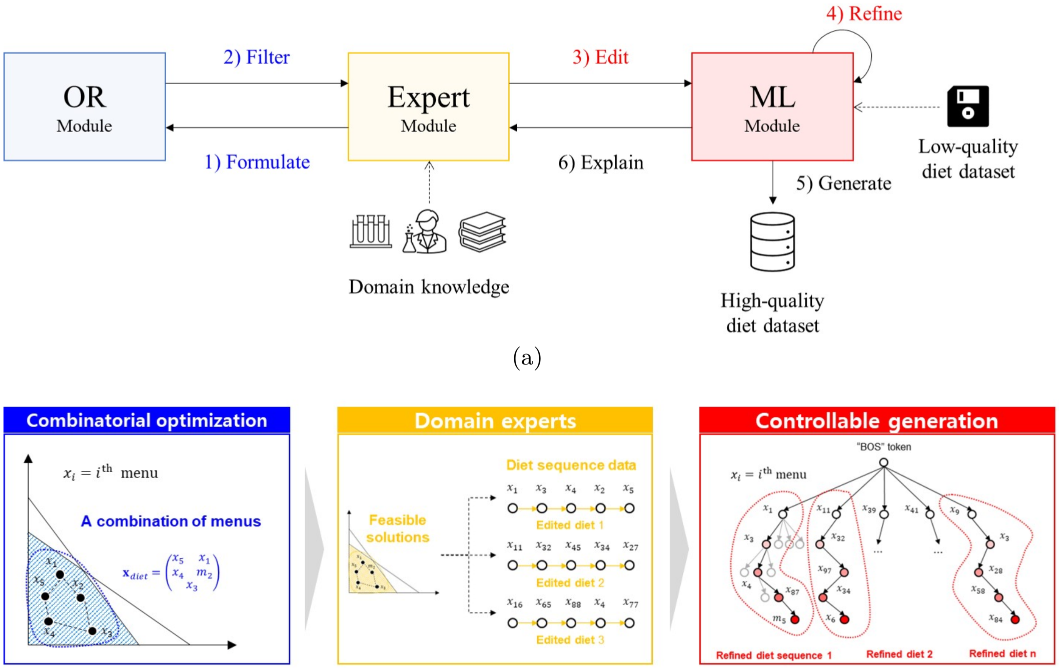
$$\sum_{t=1}^{T} \left(\mathbf{C}^{\mathrm{T}}\mathbf{X}\right)_{t} \leq C \tag{2}$$

$$\sum_{t=1}^{T} \left(\mathbf{S}^{\mathrm{T}}\mathbf{X}\right)_{kt} = T \qquad \text{for} \quad k = 1, 2, 3, ..., K \tag{3}$$

$$...$$

$\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_T] \in \{0,1\}^{M \times T}$ is the matrix representation of a single diet consisting of T menus out of l M available menus. $\mathbf{x}_t \in \{0,1\}^M$ is the menu representation of tth menu in a single diet. This is the one-hot vector of size M, the jth element of which is marked as one if the jth element is the tth menu in a diet. The remaining elements are assigned zeros. $\mathbf{A} \in \mathbb{R}^{M \times N}$ is the menu-nutrient matrix in which the value of each element is an amount of nutrient contained in one unit of the menu. N is the number of nutrients. $\mathbf{C} \in \mathbb{R}^M$ is the menu-cost vector in which the value is the unit cost of each menu. C is the upper bound of the total cost available for spending. $\mathbf{S} \in \{0,1\}^{M \times K}$ is the menu-category indicator matrix in which s_{mk} is set to one if the mth menu belongs to the category k and zero otherwise. Note that the subscript of the matrix from equation (1) to (3) indicates an indexing after inner-product operations. For emphasis, the equations are not the absolute forms of formulating the diet planning problem; rather, a dual form can be designed. One example has the objective of minimizing the total cost, given the constraint in which some amounts of the total nutrition can be achieved. Figure 3 in Appendix A.1 illustrates the concept of diet planning.

Diet planning with ML In the previous section, we described diet planning as a combinatorial optimization problem; this is the main approach used in OR communities. However, using OR approaches solely is insufficient when a problem is dynamic or comprises latent elements, but ML is an emerging approach that can help overcome the limitations of OR (see Appendix A.1). In this work, we assert that a ML-supported OR approach is the best approach to diet planning. We offer two reasons to support this view. First, diet planning is a task that requires expertise. We discovered from our interviews with diet experts that rules or practices of planning are tightly adhered to. Second, some elements are difficult to define, even among diet experts. Elements such as individual preference that include the color, flavor, or texture of menus and context that includes mealtime, food culture, or composition of the diet, are implicit information such that all possible scenarios cannot be explicitly defined. In summary, the nature of diet planning is essentially static, and various latent elements are highly implicit and difficult to describe explicitly. More specifically, we found in our previous study [35] that dietitians consider the chemistry of the menus, i.e., the composition of a diet, to be as important as meeting the nutritional requirements. We, therefore, propose a diet planning framework that addresses both explicit (nutrition) and implicit (composition) requirements. See Appendix A.4 for further details on these explicit and implicit requirements of diets. The framework consists of OR and ML parts,



(a)

(b)

Figure 2: OR-Xpert-ML framework (ORxML). Figure (a) shows that the framework consists of three modules and six steps. Steps in blue are related to explicit requirements, and steps in red are related to implicit requirements. Note that this framework is designed to have a cyclical structure that is intended to reflect the continuous interaction between experts and machines. Figure (b) describes the task of each module, e.g., combinatorial optimization and the form of the corresponding diet, e.g., a combination of menus.

each responsible for nutrition and composition; and the outputs are high-quality diet plans addressing both requirements. We elaborate on the details in the next section.

3 Methodology

To generate a high-quality diet-level dataset, we propose the ORxML framework. This framework integrates the input of experts with the capabilities of OR and ML. In particular, this section presents the overall framework and then describes each of the three modules in detail.

The overall framework As illustrated in Figure **??**, the proposed framework consists of three modules and has a cyclic structure of six repeated steps. Each of the three modules has a slightly different view of diet planning. In the OR module, diet planning is a combinatorial optimization problem to find a feasible set of menus that achieves nutritional requirements. In the expert module, the experts define diet planning as a problem that includes composition requirements and occasionally sacrifices meeting nutritional requirements to achieve desirable composition. In the ML module, diet planning is defined as the midpoint between the previous two modules. A machine is trained to control diet generation, allowing this generation to proceed as intended. We intend the algorithm to refine the diets into high-quality ones that recover the nutritional requirements obtained in the OR module while maintaining the compositional requirements provided by experts.

Six steps for the three modules The six steps specify the tasks of each module. In the first step, we **formulate** an OR model to define a searching space. The search space is the space of feasible diets defined as a set of menus that satisfy the nutritional requirements. In the second step, we **filter** the optimal solutions. An optimal solution is a combination of menus in the feasible space. The aim of this step is to filter the diets that satisfy nutrient requirements and to include these diets as candidates for the initial settings of the diet dataset. The OR module is guaranteed to find existing optimal solutions.

In the third step, the candidate diets are given to experts, e.g., dietitians, who **edit** these diets to ensure acceptability. Note that the experts are guided to edit the diets and arrange the edits into a sequential format. Such a guide is necessary because the following module, the ML module, is designed to learn the sequential patterns of menu compositions in diets. The goal of this step is to create a diet dataset with desirable compositions using the leverage of human experiences. However, manual editing is labor-intensive with the risk of biased editing or mistakes. To automate the editing task and to prevent the risk of human error, we added an ML module in the fourth step to **refine** the input diets. The goal is the compliance of compositional patterns in generated diets and the enhancement of the nutritional rewards. We designed the reward functions based on the nutritional requirements in the National Health Guideline provided by the Ministry of Health and Welfare of the Korean government. Then, based on Lee et al. [35], we developed a sequence generation model that maximizes the rewards using policy-based reinforcement learning with the REINFORCE algorithm [73]. Since reinforcement learning is built on the Markov assumption, the generative model could be based on any type of neural networks that belong to the family of RNNs, e.g., GRU and LSTM. This is powerful in learning sequential patterns and the reason that the experts were guided to edit diets in the sequence form. The ML module is trained by learning the edited diet data from the Expert module and the nutritional knowledge. Through this step, we control the candidate generated diets to be excellent in both compositional and nutritional quality.

The first two steps, i.e., formulation and filtering, are devised to consider explicit requirements; the middle two steps, editing and refinement, are introduced to consider implicit requirements. We can **generate** as many diets as is necessary in the fifth step. With a trained ML model, the process of diet generation is totally automated. In the sixth step, the learned parameters, e.g., the coefficients of the model and the attention map, of the ML model provide an **explanation** of the latent elements, implicit requirements that are unobservable to experts. Finally, these six steps can be repeated as many times as necessary such that the outcomes of the generation and explanation steps may motivate the experts to reformulate the OR model and to improve the diet editing process. See Appendix A.2 for further details of each module.

4 Evaluation

In this section, we describe the experiment settings, including the models and algorithms implemented in the OR and ML modules. In addition, we introduce three measures to evaluate the usefulness of the ORxML framework and its outcome, the diet dataset. Finally, we discuss the evaluation results.

Experiment settings For the OR module, we formulated the problem of combinatorial optimization. Then, we solved the problem using the branch and cut method [34, 44], a popular optimization algorithm in OR communities, and found optimal MIP solutions (see Appendix A.3). For the ML module, we define diet planning as a task of neural machine translation (NMT) that maps the source diets, i.e., edited diets from the expert module, into the refined target diets. Furthermore, we applied reinforcement learning (RL) to control the generative translation process such that the translated diet becomes more nutritious than the source diet [35]. All of the details of this approach are provided in Appendix A.2.

We evaluate the quality of diets generated by the modules of the ORxML framework with three measures. First, we count the number of nutrients that satisfy the nutritional standards using the Diet-Nutrition data in MIND (see Figure 2 in Supplementary material B). For the nutritional standards, we referred to the recommended dietary intake (RDI) provided by the Ministry of Health and Welfare of the Korean government (see Table 5 in the Supplementary material). We applied 15 nutritional evaluation criteria and assigned one point each time the diet satisfied a nutritional criterion. Therefore, the perfect score was 15. We named this measure the RDI score. Note that we also applied the RDI score for the constraint design and reward shaping in the OR and ML modules, respectively. Second, we calculate the ratio of mispositioned menus to evaluate the compositional quality of diets. A menu is considered mispositioned when placed as a side dish but located in the position of the main dish or vice versa. In this study, the diet has a sequence length of $T = 19$ in which each token x_t represents a menu served as a tth dish according to a rule of the dietitians' table (refer to Table 11 in the Supplementary material). This means that each menu occupies a feasible position in the diet sequence, and a diet consisting of mispositioned menus is not acceptable. Third, we performed a $\mathcal{X}^2$ homogeneity test in terms of ingredient usages. (According to the diet planning policy developed by professional dietitians, an ingredient-based diet evaluation is important. See Section A.2 and the Supplementary material A.3 for further details). $\mathcal{X}^2$ evaluates how similar the pattern of ingredient usage is between the generated diets and actual diets. Specifically, the pattern was defined based on the type and frequency of ingredients used for each meal, breakfast, lunch, and dinner. Here, we computed the co-occurrence frequency of ingredients over meals and regarded this as a homogeneity measure between the generated diets and actual diets. If the ingredients usage of each meal in a generated diet is similar to that of an actual diet, then their $\mathcal{X}^2$ value decreases. Given that usage of ingredients represents implicit compositional patterns of the flavors, colors, and textures, we compare the diets generated by each module to the actual diets using the following measure. $\mathcal{X}^2$ measures whether generated diets have the same population in terms of implicit patterns as actual ones.

Results Table 1 shows the experimental result of the diet data generated by the ORxML framework. This table shows the quality of the generated diets compared with the actual diets over the RDI score, % Mispos, and $\mathcal{X}^2$. The RDI score represents nutritional excellence of diets, and % Mispos, mispositioned menu items, and $\mathcal{X}^2$ indicate the compositional compliance with respect to the dishes and ingredients in diets.[3] The results verify that the ORxML framework succeeds in increasing diet quality. As shown in Table 1, the OR module generates diets of perfect nutrition as expected; we provide the average nutrition of generated diets and report the achievement ratio of the nutritional standards in the bracket. This is self-evident considering the characteristics of the MIP model and algorithm we used that guarantees optimal solutions unless there is no feasible one within the constraints.

However, as the % Mispos and $\mathcal{X}^2$ values denote, the compositional qualities of diets generated by the OR module are low. Note that the composition-related criteria shown in Table 1 can cover few aspects of the compositional quality of diets, and designing a metric to measure all aspects of the compositional quality is impossible, especially regarding the implicit requirements of diet planning (see Appendix A.4). Thus, we conducted a survey of 51 professional dietitians to further evaluate the compositional quality in a relatively qualitative way. The survey participants rated the compositional quality of diets generated by the OR module as low. For the expert module, the compositional quality of diets from the OR module could be enhanced by editing the diets into a more realistic form. However, the average nutritional quality declined due to the limited capability of experts to consider nutrition. The ML module recovered the nutritional quality sacrificed by the expert module; this is encouraging and as expected. In addition, the ML module outperformed the composition-related measures. That the ML module can further increase the RDI score after a sufficient training time is significant; we trained our ML model for only 40 hours[4] Additionally, as mentioned in Section 3, our framework is able to provide the experts with explanations of the compositional patterns (implicit requirements) using the attention mechanism. See Appendix A.5 for further details on the explanation of the ML module. Furthermore, in Appendix A.6, we explain the evaluation survey completed by 51 professional dietitians in detail, which is "the human evaluation of our dataset". In summary, the quality of the generated diets was validated by the experts. Furthermore, this qualitative evaluation of the ORxML framework and MIND dataset also showed the necessity of this work in creating and publishing the MIND dataset that incorporates the expertise of domain experts to perform combinatorial optimization and controllable generation.

[3] Note that the perfect RDI score is 15.

[4] with an Nvidia Quadro RTX 5000 GPU and Intel(R) Xeon(R) Gold 6136 CPU.

Table 1: Evaluation results of the diet data generated by the ORxML framework

	real diets		**OR**		**Expert**		**ML**	
RDI score (↑)	11.63		**15.00**		12.26		13.19	
% Mispos (↓)	–		0.43		0.06		**0.05**	
$\mathcal{X}^2$ (↓)	–		6.32		5.70		**3.61**	
Energy	1359.5	(68%)	1383.5	(100%)	1314.4	(62%)	1321.97	(72%)
Protein	56.16	(100%)	53.45	(100%)	54.72	(100%)	55.66	(100%)
% Carbo	0.61	(87%)	0.62	(100%)	0.61	(77%)	0.61	(81%)
% Protein	0.17	(100%)	0.15	(100%)	0.17	(98%)	0.17	(100%)
% Fat	0.21	(97%)	0.22	(100%)	0.22	(94%)	0.22	(98%)
Dietary Fiber	9.84	(21%)	17.52	(100%)	12.92	(74%)	13.08	(73%)
Calcium	592.6	(97%)	612.3	(100%)	538.8	(57%)	601.25	(94%)
Iron	9.26	(100%)	10.74	(100%)	9.47	(100%)	9.78	(94%)
Sodium	1978.5	(11%)	1517.4	(100%)	1663.7	(44%)	1620.81	(100%)
Vitamin A	445.3	(87%)	345.7	(100%)	349.7	(88%)	374.93	(100%)
Vitamin B1	1.15	(100%)	0.97	(100%)	0.96	(100%)	0.94	(78%)
Vitamin B2	1.32	(100%)	1.29	(100%)	1.19	(100%)	1.27	(100%)
Vitamin C	56.9	(69%)	55.56	(100%)	61.28	(87%)	71.93	(53%)
Linolenic	6210.9	(82%)	7407.3	(100%)	6965.5	(82%)	6796.78	(90%)
α-Linoleic	886.2	(44%)	869.3	(100%)	938.0	(61%)	925.70	(73%)
Time required	–		30 min		3 weeks ≤		40 hours	
# of diets	62		500		500		500	

5 MIND dataset

MIND is a dataset related to the diet and its constituent elements. We created this dataset based on the ORxML framework. Dietkit is the Python package that provides tools for MIND. MIND is distributed with the dietkit package and can be loaded and manipulated through dietkit.[5] There are four elements that comprise the MIND dataset: Diets, Menus, Ingredients, and Nutrition. Nutrition is the substances absorbed by our bodies through food consumption and includes protein and vitamins. Ingredients are the materials that are used to cook food menus. Menus are the end-products of foods cooked with ingredients. Diets are sequences of menus organized in terms of nutritional and compositional requirements. The detailed relationship between these elements is described in Figure 1 in the Supplementary material A. The MIND dataset is currently available in two languages, Korean and English. The contents of the dataset in each language version are the same. The ingredient data of MIND were extracted from the Standard Food Components provided by the Rural Development Administration of the Korean government. For each ingredient, the data consist of the name, its category, and the quantity of nutrients per 100 g. The dataset consists of a total of 3,036 ingredients; these are classified into 20 categories. Menu data were collected from food service management centers under the Ministry of Food and Drug Safety of the Korean government. A total of 3,238 menus were collected. We prepared the final data after processing the collected data. For instance, the names of the ingredients in the original data were unified, along with the names in the food ingredients data. A total of 527 food ingredients were used in the menu data. The menus were classified into 22 categories according to a policy that can be found in the Supplementary material. Additional information on each menu can be found in the “note” field. This information includes whether the menu is Korean or Western and whether the menu item is a main dish or side dish. Currently, we provide three kinds of diet data: “OR-generated diets”, "Experts-generated diets" and “ML-generated diets.” Each of the diet datasets is made up of 500 different diets. Each diet dataset consists of 19 menus: five breakfast menus, two morning snacks, five lunch menus, two afternoon snacks, and five dinner menus. Some diets have a slightly smaller number of menus. In this case, we included a dummy menu called “empty” to fit the data form. Detailed information about MIND and dietkit can be found in Appendix A.8 and in the Supplementary material.

[5] github.com/pki663/dietkit

6 Application of the MIND dataset

Many interesting applications of ML can be developed based on the MIND dataset. While we discuss various uses of the MIND dataset in detail in Appendix A.7, in this section we summarize some of the diet planning and DHR tasks that can be improved with ML using the MIND dataset. First, new interesting embeddings can be created with the MIND dataset for diet planning. For example, a Menu2Vec embedding can be created to represent the compositional patterns of menus in diets; similarly, an attention map can be extracted as exemplified in Appendix A.5. The "identification of alternative or complementary menus" is one of the frequent tasks that dietitians conduct in diet planning; this task is particularly important for those with food allergies. A Menu2Vec embedding extracted from the MIND dataset can be used to develop a system to recommend alternative menus to dietitians, considering their contributions to the nutritional and compositional quality of diets as well as their alternative or complementary relationships in diets. The authors are conducting such a clinical study for children with atopic dermatitis (AD) and food allergy (FA) who must restrict allergenic foods that could lead to fatal anaphylaxis. Precise diet planning and dietary healthcare are necessary to manage the growth and health of these children. In fact, this work on creating a novel high-quality diet dataset was initiated for our clinical study of children with AD and FA; actual diet data in practice were not of sufficient quality to be used to train a machine for the AI service. See Appendix A.7 for further details on these use cases. Note that a high-quality, large-scale diet dataset, such as the MIND dataset, is the basis for these use cases of diet planning and dietary healthcare that could not be conducted without ML.

The MIND dataset has limitations that involve future research issues. First, the diets in MIND are healthy "reference diets" for an unspecified majority of the population. Therefore, there is no guarantee that a user will prefer the provided diets for their meals. Second, the dataset should be extended to multinational and multicultural contexts. See the Supplementary D.2 for further details on our dataset maintenance plan. Third, integrating our MIND dataset with existing databases covering molecules and compounds is necessary [48]. This attempt will allow "precision diet for healthcare" (see Appendix A.7). Finally, a tree or graph structure could be adopted as a diet data structure, and the ML module used to synthesize high-quality diet data could be designed to learn diet data in a graph or tree structure. Nonetheless, the MIND dataset is the first high-quality resource for this research direction, and the ORxML framework can further encourage this research. We hope that future studies will take different approaches and expand the research field of “diet planning and dietary healthcare with ML.” We believe the proposed MIND dataset and the ORxML framework will contribute for such work.

7 Concluding remarks

Data construction efforts are essential for training machines that can subsequently assist in a variety of human tasks. This work creates a validated dataset to support diet-related human tasks with ML. In fact, our work has already created an impact in solving medical, economic, and social problems associated with diet planning and dietary healthcare. For example, the MIND dataset has been used by the Center for Children’s Food Service Management in South Korea beginning in the fall of 2021. This government organization is responsible for the support of daycare centers and kindergartens in Korea that cannot hire professional dietitians due to economic constraints. See Appendix A.7 for further details of these use cases. Furthermore, we did not construct the MIND dataset with manual effort only. We devised a systematic framework that integrates the capabilities of experts and machines to scientifically and efficiently create high-quality data of the complex professional diet planning task. We hope our ORxML framework can be used to inspire and promote dataset preparation methodologies and ML applications for other professional tasks. See Appendix A.7 for a detailed discussion of the value of our ORxML framework.

The authors bear all responsibility in the case of a violation of rights.

References

[1] Russell L Ackoff. The future of operational research is past. *Journal of the operational research society*, 30(2):93–104, 1979.

[2] R Amriitha, M Karpaga Meena, U Fathima Shafa, and RB Vandhana. Cd-diet: A prediction and food recommendation system for chronic diseases.

[3] Niclas Andréasson, Michael Patriksson, and Anton Evgrafov. *An introduction to continuous optimization: foundations and fundamental algorithms*. Courier Dover Publications, 2020.

[4] Jeong-Sook Bae and Kyung-Eun Lee. Nutrient consumption of children from lunch at child day care centers and kindergartens. *Journal of the Korean Society of Food Culture*, 34(6):707–718, 2019.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[6] Anita Bean. *The complete guide to sports nutrition*. Bloomsbury Publishing, 2017.

[7] Yoshua Bengio, Andrea Lodi, and Antoine Prouvost. Machine learning for combinatorial optimization: a methodological tour d'horizon. *European Journal of Operational Research*, 290(2):405–421, 2021.

[8] Sara E Benjamin-Neelon. Position of the academy of nutrition and dietetics: Benchmarks for nutrition in child care. *Journal of the Academy of Nutrition and Dietetics*, 118(7): 1291–1300, 2018.

[9] Kristin P Bennett and Emilio Parrado-Hernández. The interplay of optimization and machine learning research. *The Journal of Machine Learning Research*, 7:1265–1281, 2006.

[10] Giovanni Cammarota, Gianluca Ianiro, Anna Ahern, Carmine Carbone, Andriy Temko, Marcus J Claesson, Antonio Gasbarrini, and Giampaolo Tortora. Gut microbiome, big data and machine learning to promote precision medicine for cancer. *Nature reviews gastroenterology & hepatology*, 17(10):635–648, 2020.

[11] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[12] Benjamin L Cohen, Sally Noone, Anne Muñoz-Furlong, and Scott H Sicherer. Development of a questionnaire to measure quality of life in families with a child with food allergy. *Journal of Allergy and Clinical Immunology*, 114(5):1159–1163, 2004.

[13] George B Dantzig. Origins of the simplex method. In *A history of scientific computing*, pages 141–151. Association for Computing Machinery, 1990.

[14] George Bernard Dantzig. *Linear programming and extensions*, volume 48. Princeton university press, 1998.

[15] Enza D'Auria, Erica Pendezza, and Gian Vincenzo Zuccotti. Personalized nutrition in food allergy: Tips for clinical practice. *Frontiers in pediatrics*, 8, 2020.

[16] Arie-Willem de Leeuw, Stephan van der Zwaard, Rick van Baar, and Arno Knobbe. Personalized machine learning approach to injury monitoring in elite volleyball players. *European journal of sport science*, pages 1–10, 2021.

[17] KW DeGregory, P Kuiper, T DeSilvio, JD Pleuss, R Miller, JW Roginski, CB Fisher, D Harness, S Viswanath, SB Heymsfield, et al. A review of machine learning in obesity. *Obesity reviews*, 19(5):668–685, 2018.

[18] A DunnGalvin, BMJ De BlokFlokstra, AW Burks, AEJ Dubois, and J O'B Hourihane. Food allergy qol questionnaire for children aged 0–12 years: content, construct, and cross-cultural validity. *Clinical & Experimental Allergy*, 38(6):977–986, 2008.

[19] Eleanor Eckstein. Communication to the editor—is the "diet problem" identical to the "menu planning problem"? *Management Science*, 16(9):527–528, 1970.

[20] Aaron Ferber, Bryan Wilder, Bistra Dilkina, and Milind Tambe. Mipaal: Mixed integer program as a layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1504–1511, 2020.

[21] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

[22] Rozenn Gazan, Chloé MC Brouzes, Florent Vieux, Matthieu Maillot, Anne Lluch, and Nicole Darmon. Mathematical optimization to explore tomorrow's sustainable diets: a narrative review. *Advances in Nutrition*, 9(5):602–616, 2018.

[23] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018.

[24] Stephanie P Goldstein, Fengqing Zhang, John G Thomas, Meghan L Butryn, James D Herbert, and Evan M Forman. Application of machine learning to predict dietary lapses during weight loss. *Journal of diabetes science and technology*, 12(5):1045–1052, 2018.

[25] Neil Hill, Joanne Fallowfield, Susan Price, and Duncan Wilson. Military nutrition: maintaining health and rebuilding injured tissue. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1562):231–240, 2011.

[26] J. Hudgins, L. Bundonis, P. Tortorici, Belfer Center for Science, and International Affairs. *The Department of Defense, Artificial Intelligence, and Healthcare: Why Leveraging Data Will Improve Servicemember Health and Chart a Pathway for the Public Good.* Belfer Center for Science and International Affairs, 2020. URL `https://books.google.co.kr/books?id=eNMmzgEACAAJ`.

[27] Sarimah Ismail, Normah Kadir, Arif Kamisan Pusiran, Irina Safitri Zen, and Aqeel Khan. The importance of menu variety experience for public health sustainability at higher education institution. *Indian Journal of Public Health Research & Development*, 10(9), 2019.

[28] Min Jung, Chiehyeon Lim, Changhun Lee, Soohyeok Kim, and Jayun Kim. Human dietitians vs. artificial intelligence: Which diet design do you prefer for your children? *Journal of Allergy and Clinical Immunology*, 147(2):AB117, 2021.

[29] James E Kelley, Jr. The cutting-plane method for solving convex programs. *Journal of the society for Industrial and Applied Mathematics*, 8(4):703–712, 1960.

[30] Soowon Kim, Pamela S Haines, Anna Maria Siega-Riz, and Barry M Popkin. The diet quality index-international (dqi-i) provides an effective tool for cross-national comparison of diet quality as illustrated by china and the united states. *The Journal of nutrition*, 133(11):3476–3484, 2003.

[31] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[32] James Kotary, Ferdinando Fioretto, Pascal Van Hentenryck, and Bryan Wilder. End-to-end constrained optimization learning: A survey. *arXiv preprint arXiv:2103.16378*, 2021.

[33] Richard B Kreider, Colin D Wilborn, Lem Taylor, Bill Campbell, Anthony L Almada, Rick Collins, Mathew Cooke, Conrad P Earnest, Mike Greenwood, Douglas S Kalman, et al. Issn exercise & sport nutrition review: research & recommendations. *Journal of the international society of sports nutrition*, 7(1):1–43, 2010.

[34] Eugene L Lawler and David E Wood. Branch-and-bound methods: A survey. *Operations research*, 14(4):699–719, 1966.

[35] Changhun Lee, Soohyeok Kim, Chiehyeon Lim, Jayun Kim, Yeji Kim, and Minyoung Jung. Diet planning with machine learning: Teacher-forced reinforce for composition compliance with nutrition enhancement. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3150–3160, 2021.

[36] Pingsun Leung, Kulavit Wanitprapha, and Lynne A Quinn. A recipe-based, diet-planning modelling system. *British Journal of Nutrition*, 74(2):151–162, 1995.

[37] M Barbara E Livingstone and Alison E Black. Markers of the validity of reported energy intake. *The Journal of nutrition*, 133(3):895S–920S, 2003.

[38] Wenyin Loh and Mimi LK Tang. The epidemiology of food allergy in the global context. *International journal of environmental research and public health*, 15(9):2043, 2018.

[39] Akash Maurya, Rahul Wable, Rasika Shinde, Sebin John, Rahul Jadhav, and R Dakshayani. Chronic kidney disease prediction and recommendation of suitable diet plan by using machine learning. In *2019 International Conference on Nascent Technologies in Engineering (ICNTE)*, pages 1–4. IEEE, 2019.

[40] Vesanto Melina, Winston Craig, and Susan Levin. Position of the academy of nutrition and dietetics: vegetarian diets. *Journal of the Academy of Nutrition and Dietetics*, 116 (12):1970–1980, 2016.

[41] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[42] Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. A survey on food computing. *ACM Computing Surveys (CSUR)*, 52(5):1–36, 2019.

[43] Weiqing Min, Chunlin Liu, and Shuqiang Jiang. Towards building a food knowledge graph for internet of food. *arXiv preprint arXiv:2107.05869*, 2021.

[44] John E Mitchell. Branch-and-cut algorithms for combinatorial optimization problems. *Handbook of applied optimization*, 1:65–77, 2002.

[45] MW Murimi, M Chrisman, LK Diaz-Rios, HR McCollum, and O Mcdonald. A qualitative study on factors that affect school lunch participation: Perspectives of school food service managers and cooks. *Journal of Nutrition and Health*, 1(2):1–6, 2015.

[46] Oxford. Diet. URL `http://www.askoxford.com/concise_oed/diet_1?view=uk`.

[47] Gabrielle O'Kane. A moveable feast: Contemporary relational food cultures emerging from local food networks. *Appetite*, 105:218–231, 2016. ISSN 0195-6663. doi: https://doi.org/10.1016/j.appet.2016.05.010. URL `https://www.sciencedirect.com/science/article/pii/S0195666316301817`.

[48] Donghyeon Park, Keonwoo Kim, Seoyoon Kim, Michael Spranger, and Jaewoo Kang. Flavorgraph: a large-scale food-chemical graph for generating food representations and recommending food pairings. *Scientific reports*, 11(1):1–13, 2021.

[49] David R. Peryam. Discussion: Linear programming models for the determination of palatable human diets. *Journal of Farm Economics*, 41(2):302–305, 1959. ISSN 10711031. URL `http://www.jstor.org/stable/1235156`.

[50] J Philip Karl, Lee M Margolis, Joanne L Fallowfield, Robert B Child, Nicola M Martin, and James P McClung. Military nutrition research: contemporary issues, state of the science and future directions. *European Journal of Sport Science*, (just-accepted):1–23, 2021.

[51] Edmarlyn M Porras, Arnel C Fajardo, and Ruji P Medina. Solving dietary planning problem using particle swarm optimization with genetic operators. In *Proceedings of the 3rd international conference on machine learning and soft computing*, pages 55–59, 2019.

[52] Salman Shirvani Rad, Amirabbas Nikkhah, Mohammadmahdi Orvatinia, Hanieh-Sadat Ejtahed, Negar Sarhangi, Seyed Hamid Jamaldini, Nazli Khodayari, Hamid Reza Aghaei Meybodi, and Mandana Hasanzad. Gut microbiota: a perspective of precision medicine in endocrine disorders. *Journal of Diabetes & Metabolic Disorders*, pages 1–8, 2020.

[53] Griffin P Rodgers and Francis S Collins. Precision nutrition—the answer to "what to eat to stay healthy". *Jama*, 324(8):735–736, 2020.

[54] Zainur Romadhon, Eko Sediyono, and Catur Edi Widodo. Various implementation of collaborative filtering-based approach on recommendation systems using similarity. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, pages 179–186, 2020.

[55] Daniel A Schupack, Ruben AT Mars, Dayne H Voelker, Jithma P Abeykoon, and Purna C Kashyap. The promise of the gut microbiome as part of individualized treatment strategies. *Nature Reviews Gastroenterology & Hepatology*, pages 1–19, 2021.

[56] S Schürmann, M Kersting, and U Alexy. Vegetarian diets in children: a systematic review. *European journal of nutrition*, 56(5):1797–1817, 2017.

[57] Xianwen Shang, Yanping Li, Haiquan Xu, Qian Zhang, Ailing Liu, Songming Du, Hongwei Guo, and Guansheng Ma. Leading dietary determinants identified using machine learning techniques and a healthy diet score for changes in cardiometabolic risk factors in children: A longitudinal analysis. *Nutrition journal*, 19(1):1–16, 2020.

[58] Eun-Kyung Sin and Yeon-Kyung Lee. Menu development and evaluation through eating behavior and food preference of preschool children in day-care centers. *Journal of the Korean Society of Food Culture*, 20(1):1–14, 2005.

[59] David Sklan and Ilana Dariel. Diet planning for humans using mixed-integer linear programming. *British Journal of Nutrition*, 70(1):27–35, 1993.

[60] Paul E Smith. The diet problem revisited: A linear programming model for convex economists. *Journal of Farm Economics*, 43(3):706–712, 1961.

[61] Victor E Smith. Linear programming models for the determination of palatable human diets. *Journal of Farm Economics*, 41(2):272–283, 1959.

[62] Romeshwar Sookrah, Jaysree Devee Dhowtal, and Soulakshmee Devi Nagowah. A dash diet recommendation system for hypertensive patients using machine learning. In *2019 7th International Conference on Information and Communication Technology (ICoICT)*, pages 1–6. IEEE, 2019.

[63] A Stensgaard, Carsten Bindslev-Jensen, D Nielsen, M Munch, and Audrey DunnGalvin. Quality of life in childhood, adolescence and adult food allergy: patient and parent perspectives. *Clinical & Experimental Allergy*, 47(4):530–539, 2017.

[64] George J Stigler. The cost of subsistence. *Journal of farm economics*, 27(2):303–314, 1945.

[65] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[66] El-Ghazali Talbi. Machine learning into metaheuristics: A survey and taxonomy. *ACM Computing Surveys (CSUR)*, 54(6):1–32, 2021.

[67] De Toro-Martín, Benoit J Arsenault, Jean-Pierre Després, Marie-Claude Vohl, et al. Precision nutrition: a review of personalized nutritional approaches for the prevention and management of metabolic syndrome. *Nutrients*, 9(8):913, 2017.

[68] Thi Ngoc Trang Tran, Müslüm Atas, Alexander Felfernig, and Martin Stettinger. An overview of recommender systems in the healthy food domain. *Journal of Intelligent Information Systems*, 50(3):501–526, 2018.

[69] Thi Ngoc Trang Tran, Alexander Felfernig, Christoph Trattner, and Andreas Holzinger. Recommender systems in the healthcare domain: state-of-the-art and research issues. *Journal of Intelligent Information Systems*, 57(1):171–201, 2021.

[70] Mayumi Ueda, Mari Takahata, and Shinsuke Nakajima. Recipe recommendation method based on user's food preferences. In *Proceedings of the IADIS International Conference on e-Society*, pages 591–594, 2011.

[71] Kirill Veselkov, Guadalupe Gonzalez, Shahad Aljifri, Dieter Galea, Reza Mirnezami, Jozef Youssef, Michael Bronstein, and Ivan Laponogov. Hyperfoods: Machine intelligent mapping of cancer-beating molecules in foods. *Scientific reports*, 9(1):1–12, 2019.

[72] Paul J Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

[73] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

[74] Xiaolong Xu, Hanzhong Rong, Marcello Trovati, Mark Liptrott, and Nik Bessis. Cs-pso: chaotic particle swarm optimization algorithm for solving combinatorial optimization problems. *Soft Computing*, 22(3):783–795, 2018.

[75] Zheng-Fei Yang, Ran Xiao, Fei-Jun Luo, Qin-Lu Lin, Defang Ouyang, Jie Dong, and Wen-Bin Zeng. Food bioactive small molecule databases: Deep boosting for the study of food molecular behaviors. *Innovative Food Science & Emerging Technologies*, 66: 102499, 2020.

[76] Norlia Mohd Yusof and Shahrul Azman Mohd Noah. Semantically enhanced case adaptation for dietary menu recommendation of diabetic patients. In *Joint International Semantic Technology Conference*, pages 318–333. Springer, 2017.

[77] David Zeevi, Tal Korem, Niv Zmora, David Israeli, Daphna Rothschild, Adina Weinberger, Orly Ben-Yacov, Dar Lador, Tali Avnit-Sagi, Maya Lotan-Pompan, et al. Personalized nutrition by prediction of glycemic responses. *Cell*, 163(5):1079–1094, 2015.

[78] Simiao Zhao, Xuanyue Mao, Hanghong Lin, Hao Yin, and Peixuan Xu. Machine learning prediction for 50 anti-cancer food molecules from 968 anti-cancer drugs. *International Journal of Intelligence Science*, 10(1):1–8, 2020.

[79] Yi Zhao, Elena N Naumova, Jennifer F Bobb, Birgit Claus Henn, and Gitanjali M Singh. Joint associations of multiple dietary components with cardiovascular disease risk: A machine-learning approach. *American Journal of Epidemiology*, 2021.

A Appendix

A.1 Relationship between OR and ML

Limitations of OR As mentioned in Section 2, mathematical programming has emerged as the mainstream approach to solving diet planning problems (see Figure 3). Therefore, diet planning has been attempted exclusively using OR in academia. OR is a problem-solving approach that seeks to identify an optimal solution (or a set of optimal solutions) by directly formulating a given problem in mathematical form and applying optimization techniques such as the simplex method [13] or the cutting plane method [29]. After the OR approach became popular, researchers were able to define a problem and formulate the corresponding objective function based on their own interpretation. However, Ackoff [1] argued that OR is inherently limited. OR approaches are valid when the system of the problem is static and all of the systemic elements are tangible. However, in actual situations, the predictions, i.e.,

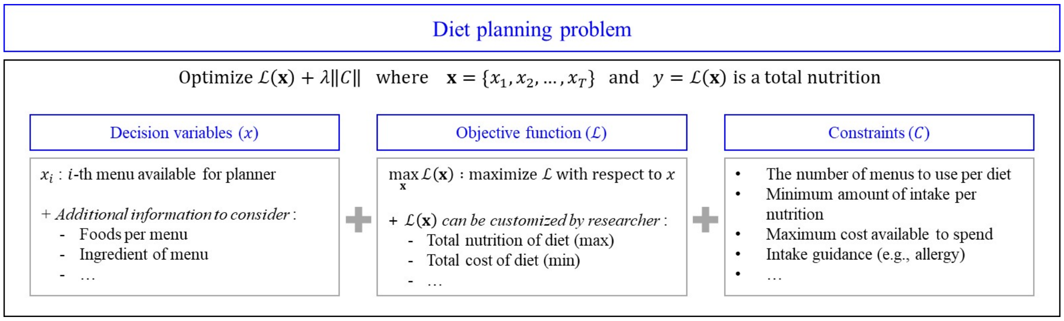

Diet planning problem
Optimize $\mathcal{L}(\mathbf{x}) + \lambda||C||$ where $\mathbf{x} = \{x_1, x_2, \dots, x_T\}$ and $y = \mathcal{L}(\mathbf{x})$ is a total nutrition
Decision variables (x)
Objective function $(\mathcal{L})$
Constraints (C)
x_i : i-th menu available for planner
+ Additional information to consider :
- Foods per menu
- Ingredient of menu
- …
$\max_{\mathbf{x}} \mathcal{L}(\mathbf{x})$: maximize $\mathcal{L}$ with respect to x
+ $\mathcal{L}(\mathbf{x})$ can be customized by researcher :
- Total nutrition of diet (max)
- Total cost of diet (min)
- …
• The number of menus to use per diet
• Minimum amount of intake per nutrition
• Maximum cost available to spend
• Intake guidance (e.g., allergy)
• …

Figure 3: Concept of diet planning

the obtained optimized solutions, typically intervene in the system dynamics. Many latent elements are unobservable. As a result, a tendency of researchers is to oversimplify actual situations using obvious, case-dependent elements to allow the application of OR approaches.

ML as a breakthrough ML has been utilized in the past few decades to overcome the limitations of the previous OR approaches and has emerged recently as a promising complementary approach. While OR seeks to optimize variables, ML focuses on parameter optimization. The following is an example of a linear optimization problem.

$$\begin{aligned} \text{minimize} \quad & \mathcal{L}_\theta(\mathbf{x}) \\ \text{subject to} \quad & f_\theta(\mathbf{x}) \geq 0 \end{aligned} \tag{4}$$

where $\mathbf{x} \in \mathbb{R}^n$ is the real-valued variable of n-dimensions and the objective function $\mathcal{L}_\theta(\mathbf{x})$ is a mean-squared-error (i.e., $||y - f_\theta(\mathbf{x})||_2$). Note that y is a target scalar variable and f_θ is a linear function that maps $\mathbf{x}$ into y using a set of coefficients parameterized by θ (i.e., $f : \mathbb{R}^n \xrightarrow{\theta} \mathbb{R}$). In the OR approach, the problem can be described as an optimal variable $\mathbf{x}^*$ that minimizes $\mathcal{L}$. The value must be obtained directly in the feasible variable space $\mathcal{X} = \{\mathbf{x}|f_\theta(\mathbf{x}) \geq 0\} \subset \mathbb{R}$ because the parameter is fixed $(\theta = \bar{\theta})$. In contrast, ML takes different approaches. In other words, the implementation of ML is somewhat indirect, but the outputs are generalizable, as depicted by the following:

$$\theta_{t+1} \leftarrow \theta_t + \frac{\partial \mathcal{L}_{\theta_t}(\mathbf{x})}{\partial \theta_t} \qquad \text{for} \quad ||\theta_{t+1} - \theta_t|| \to 0 \tag{5}$$

where the parameter θ is optimized (i.e., the machine is trained) until it converges. This is called the *steepest ascent method*. In terms of convergence, the parameter is assumed to be optimal $(\theta = \theta^*)$, and the approximated optimization can be achieved with respect to (4). This is a benefit of using ML. Moreover, whereas OR seeks to directly obtain the optimal solution $\mathbf{x}^*$ from the original problem, ML seeks to obtain an approximator f_{θ^*} and generalize the problem domain from the variable space to the parameter space. This manner of generalization allows researchers to address any problem with similar settings, [6] and the data-driven optimization over the parameter space enables them to explore the unobservable elements that constitute latent problems [9].

ML-supported OR ML is complementary to OR, and various endeavors have been pursued on the basis of this combined OR–ML perspective. These endeavors can be categorized into three classes. The first class is called a *two-stage approach* [20]. In the first stage of this approach, ML functions as a parameter estimator. In the second stage, the estimated parameters are used in OR to model the objective function and determine the optimal solutions. The second class is the *end-to-end approach* [32]. This class can be subdivided into two further types of approaches. The first type focuses on the implementation of OR methods within the ML framework by avoiding loss in the end-to-end structure. The

[6] Whereas OR performs a case-dependent modeling, ML is case-free and instead performs data-dependent modeling.

second type attempts to develop an end-to-end learning algorithm for discovering the optimal policies that are generalizable to any OR-related problems. The third class represents approaches that fall between the first and second classes. The methods in this class focus on replacing some of the OR components, e.g., initialization of the solutions and search strategies, with ML components; then the components simultaneously interact. Thus, the third class is called the *interactive approach.* The common goal of these three classes is to solve OR-related problems with the support of ML to achieve significantly better performance than pure OR approaches can obtain. Our work belongs to the third class of interactive approaches, and we refer readers to [7, 66, 9] for the details of the OR–ML relationship.

A.2 Three modules of the ORxML framework

OR module The OR module generates initial diets for use as diet data in the MIND dataset. In this module, we formulate a MIP model and optimize this model to obtain feasible solutions. The solutions are candidate diets arranged as a set, not a sequence, of menus. We formulated the model so that its objective is to maximize the number of menus to be used in each diet under the constraints of explicit requirements, such as nutrition levels, quantity limits, and substitutes or complementary relationships. The formulation method is:

$$\max_{\mathbf{X}} \sum_{m=1}^{M} \sum_{t=1}^{T} \mathbf{X}_{mt} \tag{6}$$

subject to

$$N_i^{(L)} \leq \sum_{t=1}^{T} (\mathbf{A}^{\mathrm{T}}\mathbf{X})_{it} \leq N_i^{(U)} \ , \quad \text{for} \quad i = 1, ..., 12 \tag{7}$$

$$Q_k^{(L)} \leq \sum_{t=1}^{T} (\mathbf{S}^{\mathrm{T}}\mathbf{X})_{kt} \leq Q_i^{(U)} \ , \quad \text{for} \quad k = 1, ..., 12 \tag{8}$$

$$N_{i'}^{(L)} \leq \sum_{t=1}^{T} (\mathbf{A}'^{\mathrm{T}}\mathbf{X})_{i't} \leq N_{i'}^{(U)} \ , \quad \text{for} \quad i' = 1, ..., 6 \tag{9}$$

$$Q_{k'}^{(L)} \leq \sum_{t=1}^{T} (\mathbf{S}'^{\mathrm{T}}\mathbf{X})_{k't} \leq Q_{k'}^{(U)} \ , \quad \text{for} \quad k' = 1, 2 \tag{10}$$

$$\sum_{m=1}^{M} \sum_{t=1}^{T} \mathbf{X}_{mt} = T \tag{11}$$

$$\mathbf{X}^{(p+1)} \sim \ \{\mathbf{x}^{(p+1)} | \mathbf{x} \notin \mathbf{X}^{(p)}\} \quad \text{for} \quad p = 1, 2, ..., P \tag{12}$$

where $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_T] \in \{0, 1\}^M$ is a set of T menus available out of total M menus (T = 19 and M = 3228); $N^{(L)}, N^{(U)}$ and $Q^{(L)}, Q^{(U)}$ are the lower and upper bounds of the constraints related to nutrition and quantity respectively; $\mathbf{A} \in \mathbb{R}^{3228 \times 14}$ and $\mathbf{S} \in \mathbb{R}^{3228 \times 12}$ are respectively the nutrient-menu and category-menu matrices, consisting of 14 nutrients, 3228 menus, and 12 categories.[7] Note that each category represents the group of menu items (e.g., yogurt belongs to the snack group).

We formulated two types of constraints. The equations from (7) to (8) indicate 24 main constraints, while the equations from (9) to (10) denote the eight sub-constraints. The *main constraint* is a nutritional or quantitative requirement of a single nutrient or category. The *sub-constraint* is a linear combination of main constraints; assume $\mathcal{A}$ is a feasible space constrained by each single nutrient, u and v (i.e., $A_{u\cdot}, A_{v\cdot} \in \mathcal{A}$). $\tilde{\mathcal{A}}$ is a feasible space constrained by multiple nutrients simultaneously (e.g., $A_{u'\cdot} = A_{u\cdot} + A_{v\cdot} \in \tilde{\mathcal{A}}$). Then, $\tilde{\mathcal{A}}$ is a subspace of $\mathcal{A}$ ($\tilde{\mathcal{A}} \subset \mathcal{A}$). We introduced the concept of a sub-constraint to consider a

[7] Each kth category is labelled as follows: $k = 1$ for *rice with soup*, 2 for *rice*, 3 for *soup*, 4 for *side dish*, 5 for *main side dish*, 6 for *snack*, 7 for *empty*, and so on. The *empty* category contains an *empty* menu only and is introduced to diversify the combinations of menus.

substitute or complementary effect between menus from the nutrient or category perspective. For example, equation (7) indicates that we only consider 12 nutrition constraints for 14 nutrients given that the remaining two nutritional requirements are achieved by other nutrients (see the Appendix A.3 for further details). Equations (11) and (12) are introduced to facilitate the generative process. Equation (11) sets a diet to be generated with a fixed length of T, and equation (12) functions to create a diet in bulk.[8] Diet generation is executed by running an optimization algorithm, such as the branch and cut method [34, 44]. We implemented this algorithm using a Julia-based on the Cbc–solver, an open-source program for MIP [**?**] This MIP model filters the diets that satisfy the constraints from (7) to (12).

Expert module After filtering the initial diets generated by the OR module, we recruited several professional dietitians to evaluate and adjust the initial data in the expert module. The experts edited the diets to be more acceptable in terms of implicit requirements, e.g., the composition of diets. The edit step consists of two tasks: arrangement and replacement. In the first task, experts arrange a set of menus in sequence. Such a task is quite natural considering that diets are generally perceived as a practice or rule of food consumption that follows a certain order. Furthermore, this approach is important in that the approach separates from the traditional nutrient-only view of the diet problem. This assists in overcoming the limitations that arise with OR approaches (see Appendix A.1). Moreover, this approach provides us with an opportunity to utilize a family of ML techniques for sequence data (e.g., seq2seq [65]). In the second task, experts replace some menus with their alternatives. The intent is to develop more aesthetically desirable diets by deleting inappropriate menus and inserting alternative menus. An edit was carried out rigorously according to the five standards introduced by the experts. Table 2 shows an overview of the considerations in determining whether to edit a diet. See the Supplementary material (Refer to A.3. for planning and diet data generation and the details on the considerations that the experts used in the diet editing process.

Table 2: Five standards for editing

Number	**Reason to edit**	**Decision**
1	Improper use of menus	Edit
2	Weight correction required	
3	Recipes need changes	
4	Duplicated ingredients	
5	No reason to edit	No edit

ML module With the expert module, we can treat a diet as a sequence of menus $\tau = [x_1, x_2, ..., x_T]$. In this study, the diet has a sequence length of $T = 19$, and each token x_t represents a menu served as a tth dish.[9] The 19 tokens represent a schedule of servings, and the schedule consists of three meals, e.g., breakfast, lunch, and dinner, and two snacks, e.g., a morning snack and an afternoon snack. The diet sequence of 19 or less menu items is the standardized form of daily diets for children that have been used by the Center for Children's Food Service Management in South Korea and the Ministry of Food and Drug Safety. Given that our research objective is to create a "standard" high-quality daily diet dataset, we applied this standardized form. In meal and snack servings, the menus are listed according to a writing rule of the diet table. That is, each diet sequence is defined as an array of menus that imitates a diet table aligned by a serving schedule. However, the status of diets resulting from manual editing is vulnerable to human error such as bias or inertia. As evidence, we found that the nutrition requirements guaranteed by the OR module are damaged by the editorial process (see Section 4). This implies that the presence of implicit requirements is difficult to capture and often sacrifices satisfaction of explicit requirements.

[8] During a diet generation with a total of P epochs, the menus of every p-th diet $\mathbf{X}^{(p)}$ are not involved in the $(p+1)$-th diet $\mathbf{X}^{(p+1)}$. This prevents a single unique diet from being established and creates multiple diets instead.

[9] Note that we did not count the indicative tokens, i.e., "BOS" and "EOS". These designate the begin and end of a sequence, respectively.

Therefore, we considered the necessity of the ML module to alleviate such a risk and refine the diets into high-quality ones.

The focus of the ML module is obvious. The first function is the necessity to recover the nutritional level of the diet sacrificed in the editorial process. Second, we are required to implement a machine that controls the recovery of explicit nutritional requirement while maintaining compliance with the implicit requirements. Third, the machine should be trained using edited diets. The edited diets already achieved both a desirable nutrition level and composition to some degree. In this context, we defined the task of the ML module as a controllable sequence generation with an objective function as follows:

$$\begin{aligned} \max_{\pi_\theta} J(\theta) &= \mathop{\mathbb{E}}_{x\sim\pi_\theta}\left[\sum_{t=1}^{T}\gamma^t r(x_t, x_{t+1})\right] \\ &\approx \sum_{\substack{\tau\sim\mathcal{H} \\ \hat{\tau}\sim\pi_\theta(\tau)}}\left[\sum_{t=1}^{T}\pi_\theta(x_{t+1}|x_{0:t}) r(\hat{\tau}|\tau)\right] \end{aligned} \tag{13}$$

where x_t is a tth menu token; π_θ is a diet generator parameterized by θ. In this study, we defined the diet generator as a deep neural network having a seq2seq framework [65] built on the gated recurrent unit (GRU) [11] updated by the REINFORCE algorithm [73]; the generative process is defined as consecutive predictions of $\hat{x}_t$ for $t = 1, 2, ..., T$ to generate the refinements $\hat{\tau}$; τ is an edited diet suggested from human resources $\mathcal{H}$; $\hat{\tau}$ is a refined diet generated by π_θ; $r(\cdot)$ is a reward function that returns a reward, a numerical value which measures the RDI score of diet. The RDI is shorthand for "recommended dietary intake" that is an explicit nutrition guide to follow; Note, that a reward is returned only once when the end token x_T is observed, because we can obtain an RDI in terms of a complete diet not a part of diet. Meanwhile, we defined a reward function in tricky way:

$$r(\hat{\tau}|\tau) = \frac{r(\hat{\tau})r(\tau|\hat{\tau})}{r(\tau)} = r(\hat{\tau}) \times \frac{r(\tau,\hat{\tau})}{r(\tau)r(\hat{\tau})} \approx r(\hat{\tau}) \times \rho(\tau,\hat{\tau})$$

to increase the sample diversity. In controllable sequence generation, the sample diversity means that the machine has more examples to use for training, and it enables a generative process that is more controllable with rich representations (and thereby enables a policy generalization.). As equation (13) shows, there are two independent sampling processes, of which one is for edited diets, $\tau \sim \mathcal{H}$, and the other is for refined diets, $\hat{\tau} \sim \pi_\theta(\tau)$. Then, it is obvious that we can make diverse generations with a joint sampling space such as $\mathcal{H} \bigcap \pi_\theta$. Since π_θ, i.e., a sampling distribution or policy, only changes according to rewards, we modified a reward function to force π_θ near to $\mathcal{H} \bigcap \pi_\theta$. Beginning from a conditional reward $r(\hat{\tau}|\tau)$, we derived the reward of refinements $r(\hat{\tau})$, multiplied by the reward correlation between edited and refined diets $\rho(\tau,\hat{\tau}) = \frac{r(\tau,\hat{\tau})}{r(\tau)r(\hat{\tau})}$. We defined $\rho(\tau,\hat{\tau})$ to be the ratio of menus that overlap between an edited diet and its corresponding refined diet. This forces the refined diets to become as similar as possible to the edited diets. Thereby, it helps in maintaining the implicit requirements introduced by the editorial process of humans. In summary, Equation (13) indicates that the goal of the objective function is to approximate π_θ which maximizes $J(\theta)$, and π_θ is controlled to generate $\hat{\tau}$, with a reward that is greater if it complies with a composition judged by experts. Then, the controllable diet generator is optimized using the Adam optimizer [31]:

$$\theta \leftarrow \theta - \alpha\left(-\nabla_\theta J(\theta)\right) = \theta + \alpha\nabla_\theta J(\theta) \tag{14}$$

where

$$\begin{aligned} \nabla_\theta J(\theta) &= \sum_{\substack{\tau\sim\mathcal{H} \\ \hat{\tau}\sim\pi_\theta(\tau)}}\left[\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(x_{t+1}|x_{0:t}) r(\hat{\tau}|\tau)\right] \\ &= \mathop{\mathbb{E}}_{\substack{\tau\sim\mathcal{H} \\ \hat{\tau}\sim\pi_\theta(\tau)}}\left[r(\hat{\tau}|\tau)\nabla_\theta \log \pi_\theta(\tau)\right]. \end{aligned} \tag{15}$$

To implement and execute equation (15), we applied the *teacher-forced REINFORCE* algorithm (TFR) suggested by Lee et al. [35], which is one of the most recent papers that address the diet planning problem with ML leverage.

A.3 Details of MIP formulation in the OR module

In this section, we introduce some techniques and more details concerning the OR module. For a typical OR problem, a unique solution is found with an objective function that minimizes or maximizes values such as cost. However, unlike the usual problem, we needed to generate different diets satisfying the conditions and used a slightly different method to accomplish this. As shown in Equation (6), we used an objective function that maximizes the number of menus used in the diet, but we have fixed the number of menus in the diet to 19 as a constraint. We prevent using the menu of the previous diet, as shown in Equation (12), to avoid the repeated usage of a specific menu. Specifically, we calculated the appearance frequency of menus in generated diets and selected multiple menus to remove at the next generative process. The number of the menus to remove can be customized, and we removed ten menus in every epoch. In this way, we could continuously generate a new diet that satisfies nutritional and basic composition constraints. A more detailed description of the two equations (9) and (10) follows. As shown in Table 5 in the Supplementary material, there are three requirements for the percentage of carbohydrates, protein, and fat. The reason for this is that the proportion of each of these nutrients is affected by the number of calories provided by each of these nutrients. Thus, in order to make a linear inequality with a predefined value, we reorganized our formula below into a different form:

$$\begin{aligned}
0.55 \leq & \frac{4\sum_{t=1}^{T}\left(carbohydrate^{\mathrm{T}}\mathbf{X}\right)_t}{\sum_{t=1}^{T}\left(calorie^{\mathrm{T}}\mathbf{X}\right)_t} \leq 0.65 \\
0.07 \leq & \frac{4\sum_{t=1}^{T}\left(protein^{\mathrm{T}}\mathbf{X}\right)_t}{\sum_{t=1}^{T}\left(calorie^{\mathrm{T}}\mathbf{X}\right)_t} \leq 0.20 \\
0.15 \leq & \frac{9\sum_{t=1}^{T}\left(fat^{\mathrm{T}}\mathbf{X}\right)_t}{\sum_{t=1}^{T}\left(calorie^{\mathrm{T}}\mathbf{X}\right)_t} \leq 0.30
\end{aligned}$$

By rearranging the equations, we converted three inequalities to six inequalities, as shown below:

$$\begin{aligned}
0 \leq & \sum_{t=1}^{T}\left(\left(4carbohydrate - 0.55calorie\right)^{\mathrm{T}}\mathbf{X}\right)_t \\
& \sum_{t=1}^{T}\left(\left(4carbohydrate - 0.65calorie\right)^{\mathrm{T}}\mathbf{X}\right)_t \leq 0 \\
0 \leq & \sum_{t=1}^{T}\left(\left(4protein - 0.07calorie\right)^{\mathrm{T}}\mathbf{X}\right)_t \\
& \sum_{t=1}^{T}\left(\left(4protein - 0.20calorie\right)^{\mathrm{T}}\mathbf{X}\right)_t \leq 0 \\
0 \leq & \sum_{t=1}^{T}\left(\left(9fat - 0.15calorie\right)^{\mathrm{T}}\mathbf{X}\right)_t \\
& \sum_{t=1}^{T}\left(\left(9fat - 0.30calorie\right)^{\mathrm{T}}\mathbf{X}\right)_t \leq 0
\end{aligned}$$

where a coefficient matrix $\mathbf{A}'_{i'j} \in \mathbb{R}^{3228\times 6}$ is the subspace of matrix $\mathbf{A}$, and the six columns of $\mathbf{A}'$ is the stack of coefficient vectors in the six inequality constraints above. Note that the coefficients denoted as the name of each nutrient is a vector of length M whose each element represents a value of the corresponding nutrient.

Since there is a combo menu that can substitutes for rice and soup at once, we made an equation (16). In this case, the rice and soup should always be used at the same time so that we added constraints (17) to fix the number of rice and soup equal to each other within a single diet. These two equations are reflected in Equation (10) in the form of $\mathbf{S}'_{k'j} \in \mathbb{R}^{3228\times 2}$, the subspace of matrix **S**.

$$\sum_{t=1}^{T} \left((2ricewithsoup + rice + soup)^{\mathrm{T}} \mathbf{X} \right)_t = 6 \tag{16}$$

$$\sum_{t=1}^{T} \left((rice - soup)^{\mathrm{T}} \mathbf{X} \right)_t = 0 \tag{17}$$

A.4 Details of the requirements and the difficulties of diet planning

Explicit and implicit criteria of diet planning Many countries define nutritional standards that are tailored to the characteristics of their own culture. Such standards are established based on studies of dietary patterns and citizen health status. The goal of establishing such standards is to prevent excesses or deficiencies of any dietary nutrient. This is accomplished by encouraging the intake of nutrients that are generally consumed in insufficient quantities. Equally, the standards are designed to strictly limit the intake of those that are consumed in excess of nutritional requirements, ultimately helping the citizens lead healthier lives. Thus, one of the responsibilities and missions of dietitians is to plan diets that comply with both cultural and nutritional standards. However, in practice, difficulty arises in planning diets based on these nutritional standards. The reason for this is that individuals do not just intake individual nutrients; our foods contain a variety of nutrients that must be balanced as a whole in our diets. Meeting nutritional requirements would be relatively easy if individuals consumed nutritional supplements with limited quantities of food. However, consumption patterns do not allow this; the combinations of nutrient quantities in individual diets are complex when considering the variety of foods consumed. Because of the difficulty in balancing nutritional intake, dietitians plan diets using a method called a "food guide." The food guide classifies types of foods into groups based on their nutritional characteristics; those having similar nutrients are classified in the same group. Based on these food groups, grains, meat/fish/eggs/beans, vegetables, fruits, milk/dairy products, and fats/oils, dietitians generate diets. These diets are designed to meet nutritional standards by determining the frequency of consumption of items from each food group. This was the method used in developing the 2019 guideline for diet planning provided by the Center for Children's Food Service Management.

However, diets generated in this manner only roughly meet nutritional standards. Upon close examination, diets planned this way were found to result in a deficiency or excess of specific nutrients when compared to the Korean Dietary Reference Intakes. Furthermore, diets used in schools or food services for childcare centers only manage the children's nutrient intake during their stay at the institutions [4]. Children that do not have access to such food services provided by the institutions, for example, those homebound due to COVID-19-induced difficulties in face-to-face education or those with food allergies, are without these planned diets. While parents generally attempt to serve their children balanced diets, the difficulties faced by professional dieticians in balancing diets are compounded for parents [27].

The diet composition, the harmony among menus, is the most important factor for dietitians in designing and planning diets [58]. In diets, menus are considered to be in harmony if various textures, colors, and food groups are represented without significant overlap [70]. The foods provided in a diet should complement or enhance one another in taste, smell, texture, and nutritional value. The ability to determine the nutritional component is solely based on the professional background knowledge of the dietitian planning the diet, and dietitians plan diets based on their estimation of the nutritional content of different foods. Dieticians' personal dietary patterns, preferences, and tendencies may be reflected in their diet planning. Knowledge of the exact nutritional value of every ingredient is not possible. This is the primary reason that "diversity" is used and emphasized in maintaining harmony when planning diets[30]. The use of ingredients from various food groups and colors minimizes the risk of an excess or deficiency of any one nutrient, and consumption of a variety of foods results in intake of a variety of nutrients.

However, composition satisfaction can limit the provision of diets with high nutritional qualities as the goals of satisfying the nutritional and compositional standards pose restrictions on each other. Efforts to satisfy more standards limit the use of various ingredients; consequently, a pattern may arise in which a certain nutritional standard can only be satisfied by including a specific food that contains a large amount of the corresponding nutrient. Adding variety to the diet composition, therefore, increases the difficulty in satisfying certain standards. While the harmony found in menus composed by dietitians can provide variety and consumer satisfaction, these diets, planned based on estimated nutrient content, may be limited in their ability to satisfy nutritional standards.

Values and implications of the MIND dataset for nutrition and healthcare Given the requirements and difficulties of diet planning, dietitians and their planned diets are more affected by the composition of the diet and the preferences of the recipients than by the nutritional standards. Although dietitians are aware that the diets should satisfy nutritional standards, there are often no options but to plan diets based on these other factors, even if the resulting diets do not satisfy the standards. Dieticians know this, and artificial intelligence (AI) may be able to solve this. AI offers the possibility of diet provision that satisfies nutritional and compositional standards. In additional to providing a nutritionally balanced diet, these diets can introduce a variety of nutrient-supplying foods, avoid overlaps of recipes or ingredients, maintain harmony in the tastes/colors/forms of foods, and fulfill consumer preference requirements. Dieticians would be able to redirect their efforts and focus on other job requirements such as hygiene or nutritional education. The positive effect that AI-generated diets would have on the growth and health of infants is especially important as parents would be able to provide higher-quality diets that fit the nutritional standards every day.

The MIND dataset is significant in that the dataset was planned through the coordination of AI systems and human dietitians with the goal of providing diets that satisfy both nutritional standards and harmonious composition requirements. With an optimization model, the draft data were developed to fulfill most of the nutritional requirements for children aged 3–5 years. Attempting to satisfy all of the nutritional standards generally limits the range of ingredients or foods that can be used. However, despite such difficulties, this MIND dataset work demonstrated the ability to create harmonious compositions while incorporating food variety. The MIND dataset will help many dietitians plan diets that maintain quality in both nutritional and aesthetic composition.

A.5 Explanations of the ML module in the ORxML framework

As mentioned in A.2, we developed an ML model based on the seq2seq framework with GRU units. GRU is in the family of recurrent neural networks (RNNs) that are widely used to address sequential information using backpropagation through time (BPTT) [72]. However, RNN-based models with BPTT have a chronic problem of an information bottleneck that causes a memory loss, and an attention mechanism [5] was proposed to overcome this problem. Attention is a technique that makes a model recognize the parts of the sequence on which to concentrate, and the attention map allows visualization of the distribution of focuses of attention.

Figure A.5 shows the attention maps extracted from the experiment in Section 3. As shown, these maps explain the menus on which to focus for the improvement of nutrition. Here, the x-axis is the input, an edited diet, and the y-axis is a refined diet, the output. The highlighted cells provide explanations regarding the menus that should be considered as candidates for replacement. Note that the highlight implies that the replacement of menus in that cell will recover the nutrition of the diet. Using this method, the experts can expand and advance their knowledge required to improve the diets being designed and to reformulate the OR module if necessary.

A.6 Details on the evaluation of the MIND and the ORxML framework

Expert evaluation As of August 2021, the evaluation experiment in Section 4 is the latest experiment, and the MIND dataset published in this work involves the outcomes from this

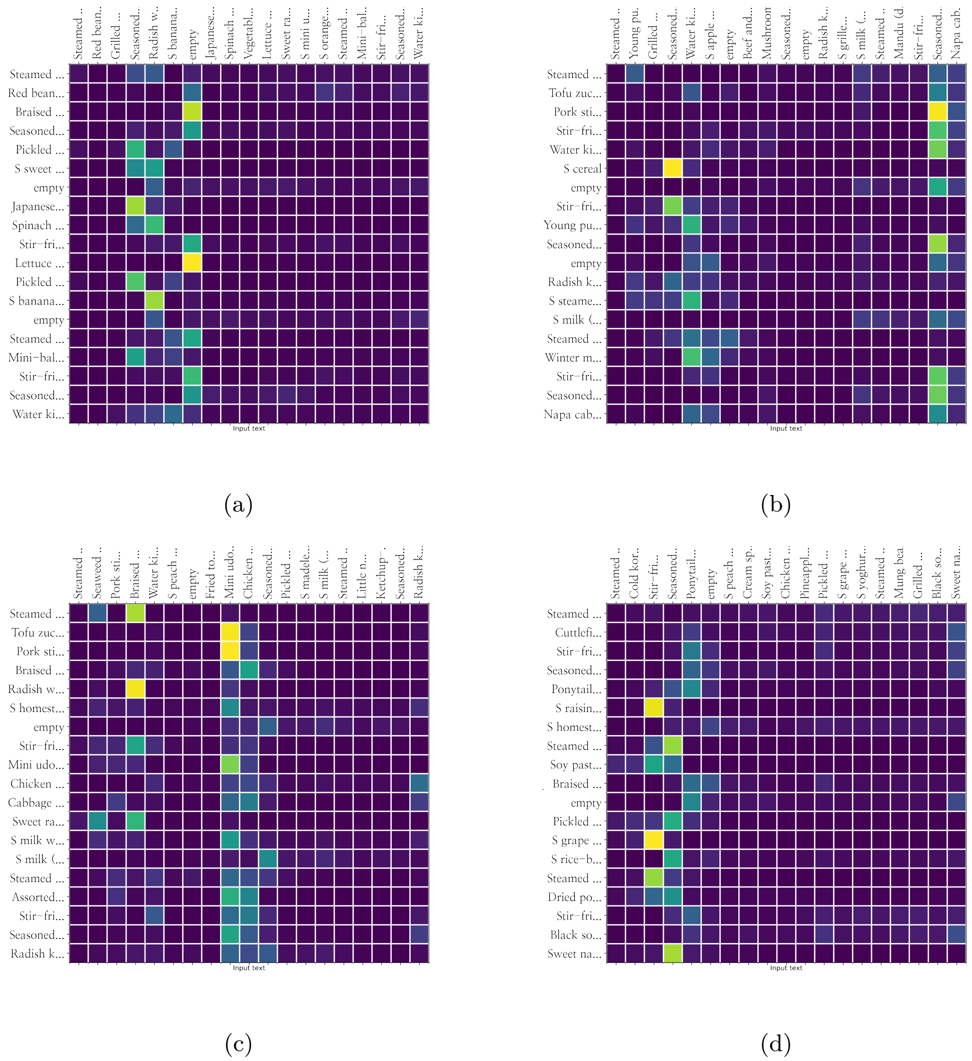


Figure 4: Attention maps of the ML module: An edited diet sequence (i.e., an input) provided by the expert module is marked on the x-axis while the y-axis denotes its counterpart, a refined diet sequence (i.e., an output), generated by the ML module.

experiment. These outcomes are the diet data from the OR module, Expert module, and ML module. Additionally, we conducted previous rounds of the evaluation experiment of the MIND dataset development process. This included the conduct of a survey of 51 professional dietitians with an average job experience of 4.85 years (range of 1 to 12 years). Although the nature of an expert survey could be qualitative in nature, the purpose of our survey was to evaluate the compositional quality of the diets synthesized by the ML module. This was required because the design of a metric to measure all aspects of the compositional quality of diets, the implicit requirements of diet planning, is impossible (see Appendix A.4). Note that the composition-related criteria shown in Table 1 in Section 4 covers few aspects of the compositional quality.

The 51 professional dietitians evaluated the 60 diets designed by the OR module, OR+Expert modules, and OR+Expert+ML modules. Table 4 shows results regarding the evaluation criteria described in Table 3 [8]. The results suggest three issues. First, experts may be biased primarily toward composition satisfaction in dietary evaluation and then apply this bias when evaluating other aspects such as the nutritional content and overall satisfaction. For example, the experts did not give a high nutrition score to the nutritionally perfect

Table 3: Form of survey

Section	No.	Evaluation criteria	Scoring
Nutrition	1.1	Does this diet satisfy the nutrition standards in terms of calories?	Integer between 1 and 5
	1.2	Does this diet satisfy the nutrition standards in terms of a balance of carbohydrate, protein, and fat?	
	1.3	Does this diet satisfy the nutrition standards in terms of vitamins and minerals?	
Composition	2.1	Is this diet balanced in terms of the food groups?	
	2.2	Does this diet harmonize in terms of the menus in meals?	
	2.3	Is this diet balanced in terms of the cooking methods?	
	2.4	Does this diet harmonize in terms of the menus in snacks?	
Overall reliability	3.1	Do you think this diet is suitable for a real service?	
	3.2	Do you think this diet was planned by a professional dietitian? (Turing test)	Yes (1) or No (0)

diets generated by the OR module. This was perplexing. Second, following this example, most of the experts were not capable of precisely evaluating the nutritional quality of diets. As shown in Section 4, the OR module and the OR+Expert+ML modules should be superior to the OR+Expert modules in nutritional excellence. However, Table 4 shows that the experts could not evaluate the nutrition of diets; this may well be due to limited calculation abilities and the biases mentioned above. Finally, although the OR+Expert+ML generated the diets to be compositionally less adequate than the human-designed diets, this result is natural because any machine-generated diet was new and unfamiliar to the experts. Also, the experts were biased by the human-designed diets publicly disclosed as a reference database by the government. The important question is whether the composition of machine-generated diets is acceptable for actual food provision. As shown with the reliability score in Table 4, we received feedback from the survey participants that the diets generated by the OR+Expert+ML modules can be used in practice. In addition, as shown in Section 4, the OR+Expert+ML modules demonstrate better performance than experts in terms of ensuring the nutritional quality of diets.

In summary, the experts considered composition compliance as the most important factor in evaluating diets in addition to being incapable of accurately evaluating the nutritional quality of diets. This lack of capability confirms the motivation of this work, to create and publish the MIND dataset using combinatorial optimization and controllable generation as no high-quality diet-level dataset exists at present due to the extreme complexity of diet planning. The survey findings confirm the necessity of the OR and ML modules in our ORxML framework, and the importance of composition compliance confirms the irreplaceable role of domain experts, professional dietitians. These experts are required for the consideration of the implicit requirements of diet planning and data generation, thereby confirming the necessity of the expert module in our ORxML framework.

Table 4: Result of survey

	Score of the evaluation criteria								
Questions	1.1	1.2	1.3	2.1	2.2	2.3	2.4	3.1	3.2
Real	4.38	4.15	3.97	3.96	3.87	3.85	3.70	3.62	0.67
OR	3.75	3.12	3.52	3.25	2.56	3.17	2.38	2.19	0.15
Expert	4.30	4.01	4.03	4.04	3.81	3.93	3.83	3.61	0.68
ML	4.26	3.92	3.80	3.80	3.61	3.82	3.39	3.29	0.55

A.7 Use of the MIND dataset and the implications of the ORxML framework

MIND for ML applications in dietary healthcare As mentioned in the introduction, a balanced diet is important for all. This includes children and senior citizens, the well and the sick . Thus, diet planning has emerged as a core part of healthcare research in a variety of fields including food technology [42], nutrition management [15], clinical medicine [77], sports science [6, 33], and military nutrition [50, 25]. The MIND dataset can be used in all of these fields.

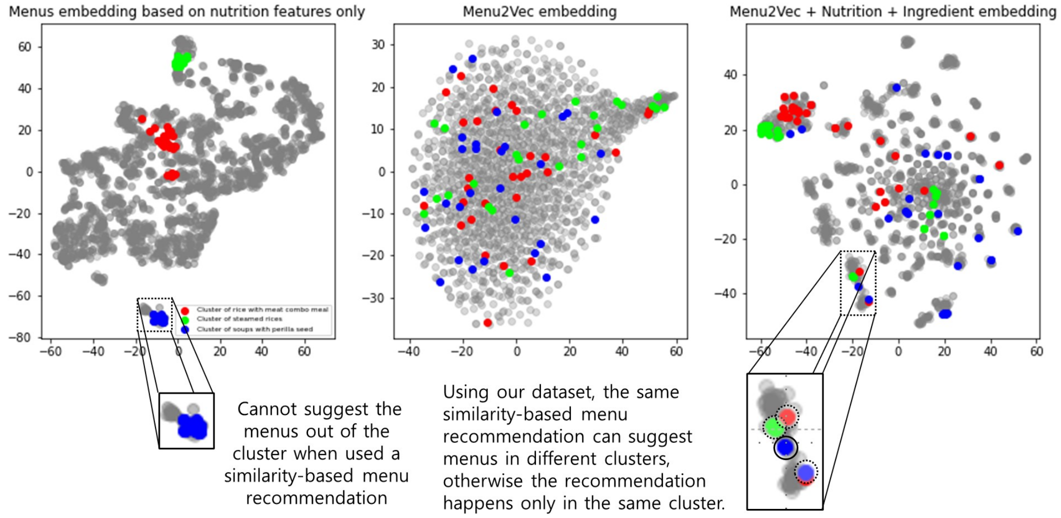
Using dietary data in the MIND dataset, ML can design diets that counter disease-related factors [77, 39, 62, 2] or to identify dietary factors that contribute to the strengthening of physical abilities and improvement of metabolic controls that are important in sports and military contexts [26, 16]. MIND, a large-volume dataset accessible to the public, is the first benchmark dataset for ML-based dietary healthcare studies.

Our work has already created an impact. The Center for Children's Food Service Management in South Korea will use our outcomes beginning in the fall of 2021. In addition, a startup company in South Korea is using this dataset for the development of a gut microbiome-personalized diet recommendation AI system for children with atopic diseases and food allergies. This service will be distributed under the support from the Ministry of Science and ICT. The authors have also started a government project to use this dataset for the development of a precision diet service system in which the users, e.g., dietitians and physicians, can work with the application interactively for precise diet planning for recipients. This service will be distributed under the support from the Ministry of Food and Drug Safety and the National Research Foundation of Korea. The following paragraphs more specifically illustrate MIND dataset utilization for ML-based dietary healthcare studies and applications.

Task 1: Complementary healthy menu recommendation Typically, people spend much of their day away from home and consume more than half of their daily energy intake away from home. Therefore, identifying complementary in-home and out-of-home menus is important. This is particularly important for growing children. Because of the increase in the number of working mothers in developed countries, the number of children attending daycare centers is significantly large. The participation rate in Korean daycare centers increased from 51.7% - 69.2% in 2010 to 53.8% - 88.8% in 2017. Therefore, parents need to be able to produce home menus for their children that nutritionally complement the diets provided in daycare centers. However, as mentioned in the introduction and Appendix A.4, this task is difficult because of the required nutrient and growth knowledge and the high complexity of design associated with large numbers of food items.

Using the MIND dataset, an application can be trained to generate healthy menu recommendations inside the home, outputs from inference, complementary to the menus consumed outside the home, inputs for inference. For example, an incomplete sequence of morning snacks, lunch meals, and afternoon snacks can be input to the algorithm, and the outputs will be a complete daily diet sequence that adds the complementary healthy menus for breakfast and dinner. We will distribute the MIND dataset to parents in South Korea through the Center for Children's Food Service Management in South Korea beginning in the fall of 2021. In addition, beginning in June 2021, we started a government project to use this dataset for the development of a precision diet service system in which the users, e.g., parents and pediatricians, can interactively work with the application for precise diet planning for children, including complementary healthy menu identification. This service will be distributed under the support from the Ministry of Food and Drug Safety and the National Research Foundation Korea.

Task 2: Allergy-free safe diet planning Vegetarians and food-allergic patients need to prepare diets customized to their unique needs [56]. Recently, there is an increase in the number of children with food allergies [38], and one of the most difficult tasks in daycare centers is to provide special care to these children. For these children, dietitians consider food allergen elimination in the meal and prepare alternative food menus tailored to meet nutritional needs in the absence of potential allergens. In addition to menu considerations, care needs to be taken to avoid cross-contamination during the cooking process. In the case of vegetarian children and adolescents, some vegetarian diets may be low in specific nutrients,



Cannot suggest the menus out of the cluster when used a similarity-based menu recommendation

Using our dataset, the same similarity-based menu recommendation can suggest menus in different clusters, otherwise the recommendation happens only in the same cluster.

Figure 5: Menu embedding using the MIND dataset and the visualization of T-SNE map.

such as calcium and vitamin B12, and their inclusion is needed to ensure proper growth and development [40]. AI is expected to reduce the time- and resource-consuming processes related to diet design and clinical nutrition.

Using the MIND dataset, AI can be trained for safe, allergen-free diet planning. A controllable generation algorithm that is similar to the ML module in the ORxML framework can be designed using reward-shaping for explicit requirements to reflect the needs of food-allergic patients. This algorithm would give negative rewards for restricted ingredients and nutrition and give positive rewards for recommended ingredients and nutrition. A startup company in South Korea has already begun to develop such an algorithm for a gut microbiome-personalized diet recommendation AI system for children with atopic diseases and food allergies. In this service, parents request a gut microbiome examination for their children, and the service identifies the type and conditions of the child's microbiome. Then, the controllable generation algorithms that are pre-trained for specific types of children's microbiome conditions are used to generate safe, allergen-free diets for the children in a customized manner. This service will be distributed under the support of the Ministry of Science and ICT. As pioneered by Zeevi et al. (2015) [77], this kind of precision diet service with ML will have considerable value in the prevention and management of chronic diseases.

Task 3: Menu2Vec embedding for menu recommendation in diet planning and dietary healthcare Figure 5 shows that new interesting embeddings can be created with the MIND dataset for diet planning and DHR. In Figure 5, the relative position among menus in the T-SNE map changes according to different embeddings. The Menu2Vec embedding represents the compositional patterns of menus in diets, and the Menu2Vec embedding is concatenated with the nutrition features. To obtain the Menu2Vec embedding, we applied the Word2Vec [41] to the generated diets. Each point in the T-SNE map is a menu, and the color indicates the cluster to which the menu belongs. Cluster assignment was determined based on the affinity propagation [21] with the Euclidean distance in terms of nutrition. As such, each of the three figures illustrates a cluster distribution of menus that are identified with the original nutrition features only.

The left figure shows that menus in the same cluster are distributed near one another. This implies that the menus of the same cluster are positioned so close in the embedding space that the choice of alternative menus, i.e., the choice of menus in different clusters, is strictly limited. However, the central figure shows that the menus of generated diets are randomly distributed regardless of the nutrition cluster. The generated diets always guarantee high nutritional quality (see Table 1), but menu embedding is not solely dependent on nutritional criteria. Therefore, simply replacing a menu with its nutritional counterpart, i.e., replacing with substitutes in the same cluster, the usual practice of dietitians, does not always provide a nutritionally better diet. Last, the right figure shows a visual demonstration of possible scenarios in which the dietitians use the MIND dataset for actual applications. For example, some dietitians design diets for patients suffering from "*perilla seed*" allergy, a common allergy usually discovered when the patient first consumes a perilla oil. These dietitians can slightly modify a reference diet from the MIND dataset, replacing a menu that contains perilla seeds with its similar, perilla-free substitutes. To support dietitians, we can develop a menu recommender system that suggests perilla-free menus by calculating the similarity between all menus. The similarity-based menu recommendation has been addressed in many previous studies [76, 68, 54, 69]. However, menus with nutritional value are likely to have similar ingredients. Therefore, a similarity-based recommendation may fail to suggest a perilla-free substitute that maintains a similar nutritional level. In this context, the MIND dataset is potentially valuable. Specifically, if a menu recommender system is trained on our dataset and calculates a similarity based on a Menu2Vec embedding, then the system can successfully suggest perilla-free menus by simply executing a similarity-based recommendation that maintains the same quality of nutrition.

Table 5 shows the result of a similarity-based recommendation for perilla-free menus. The *perilla seeds* were assumed to be an allergy-causing ingredient and *seasoned dried radish leaves and perilla seed* to be a target menu. The similarity was computed based on the embedding vector of nutritional feature and the concatenated embedding vector of Menu2Vec + Nutrition + Ingredient. Note that we purposely concatenated the ingredient feature to degenerate a performance of the recommender system; if the ingredient feature is concatenated, then the top-ranked recommendations will always include perilla-related ingredients. The similarity-based recommendation will be likely to suggest a menu that still contains the allergen. Concatenated embedding would obviously have a better recommendation. As shown in the table, the recommendation made on the basis of nutrition features alone completely failed; not only did the suggested menus still contain perilla seeds, but the menu category also changed. (The target menu belongs to the side dish category, but one of the suggested menus belongs to the soup category.) To force a recommender system to not suggest the menus with the allergen, we added a filtering step prior to similarity calculation. In the filtering step, the menus having allergens were explicitly removed. Nonetheless, the suggested menus were not satisfactory in that the menus did not belong to the same category of target menu. In contrast, the recommendation based on the MIND dataset achieved the exclusion of the allergen perfectly, and most of the suggested menus were selected from the same category of target menu, i.e., side dish. This is evidence that the menu recommender system naturally considers nutrition- and composition-related factors simultaneously when using the MIND dataset. Therefore, the MIND dataset has a variety of potential uses in healthcare applications.

Table 5: Allergeny-free menu recommendation based on MIND dataset

Embedding	Nutrition feature only	Menu2Vec + Nutrition + Ingredient
Allergen ingredient	*Perilla seed*	
Target menu	*Seasoned dried radish leaves and perilla seed* (side-dish)	
Suggested menu	Recommendation only	
	Seasoned sweet potato stem and perilla seed	*Stir-fried dried radish leaves*
	Seasoned salad with perilla seeds and cucumber	*Seasoned salad with dried radish leaves in soy paste*
	Mushroom perilla seed soup	*Bean powder dried radish leaves soup*
	Filtering and Recommendation	
	Seasoned daikon	(None)
	Frozen and dried pollack soup	(None)
	Green onion daikon soup	(None)

Task 4: Clinical studies with the MIND dataset The MIND dataset is being used for a clinical study of children with atopic dermatitis (AD) and food allergy (FA). Restriction of allergenic foods that may lead to fatal anaphylaxis is required for these children. The parents are burdened with having to limit the children's participation in social activities, e.g., camps and parties, and exposure to restaurants due to the potential of accidental ingestion of the allergenic food [63]. The children need to consume only foods prepared at home. Thus, precise diet planning and dietary healthcare are necessary to ensure the growth, development, and health of these children. The authors are working to develop an AI service application for this purpose. This work on creating a novel high-quality diet dataset was initiated for our clinical study of children with AD and FA (IRB number: KUGH IRB No. 2021-09-019). The available, practical dietary data were of insufficient quality to be used in the training of a machine for the AI service application.

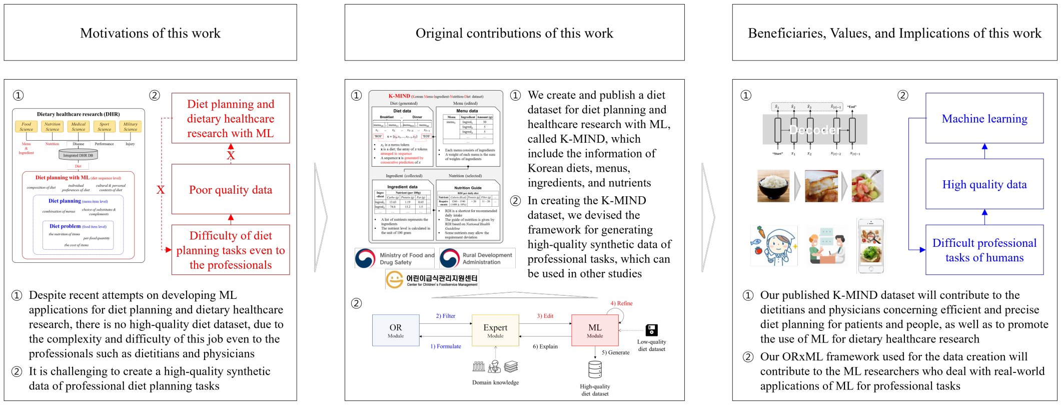
To the best of our knowledge, this is the first clinical study to use and test the utility of an AI application for diet-level planning and healthcare for patients. In this clinical study, the quantifiable health measures include: (1) the Food Allergy Quality of Life-Parental Burden Questionnaire, (2) the Food Allergy Quality Of Life Questionnaire-Parent Form, (3) Food Allergy Independent Measure, (4) nutrient assessment using 24-hour dietary recall, (5) food frequency questionnaires, (6) growth status assessments, and (7) other life satisfaction indices. These measures are validated measures in clinical studies of patients requiring food restrictions[12][18][37]. Using these measures, our clinical study will evaluate the utility of our machine for diet-level planning and healthcare in the AI service user group and the control group. In addition, we have collected gut microbiome data from children with AD and healthy children to analyze the diets-microbiota association (IRB number: KUGH IRB No. 2020-11-025-016). The gut microbiome is rapidly becoming an important factor for precision medicine in cancer [10], metabolic diseases [52], autoimmune or inflammatory diseases, and allergic diseases [55]. Diets-microbiota data are essential for precision medicine, and machine learning contributes to this approach for medical recommendations [10]. In our ongoing clinical study, diets are precisely recommended to patients with AD after analyzing their diets-microbiota associations.

Note that a high-quality large-scale diet dataset is the basis for such clinical studies on diet planning and dietary healthcare with ML. Previous to our MIND dataset, these could not be conducted. Our MIND dataset is dedicated to diet planning and dietary healthcare for allergic diseases and for chronic diseases such as diabetes mellitus, hypertension, obesity, celiac disease, gastrointestinal cancer, liver cirrhosis, and chronic kidney failure. Although there exist digital healthcare services for obesity, existing services do not tailor support to individual needs. For diabetes mellitus patients, the proportions of carbohydrate, protein, and fat must be considered carefully in diet planning, but compositional compliance is still important. For patients with chronic kidney diseases, diets should be composed of menus with decreased potassium, phosphorus, and calcium. The needs and requirements of diet planning and dietary healthcare for chronic diseases are all different, and design of quality diets that are acceptable to patients is always important. The MIND dataset is the first high-quality resource for this research direction, and the ORxML framework can further encourage this research. Physicians and dietitians have a nutrition- and menu-level dietary management education. Our work can contribute to expanding patient dietary management to the diet level.

Toward precision diet for healthcare Previous studies have also created and used representations of food-related datasets for different purposes. For example, the FlavorGraph is used to create food representations and food pairings recommendations [48], and food knowledge graphs can be used for personalized dietary recommendations and food production [43]. These studies analyzed representations at the menu, ingredient, and nutrition levels; our MIND dataset can be used to analyze representations at the diet, menu, ingredient, and nutrition levels. The addition of the diet level representation is important because of its usefulness in considering the compositional patterns of menus in diets, the identification of feasible menus, diet harmony, and the complementary nature of diets with existing food consumption habits. Our work should be connected and integrated with existing datasets with the goal of establishing a healthcare "precision diet". For example, food bioactive small molecule databases (FBSMDs) [75] provide valuable information at the levels of molecular behavior and molecular nutrition and can be used for drugs and health products. Note that precision diet is a diet-level approach for precision nutrition, a new branch of precision medicine, which aims to understand the health effects of the complex interplay among genetics, the microbiome, antibiotic and probiotic use, metabolism, food environment, and physical activity. Economic, social, and other behavioral characteristics are also included in this analysis [67, 53]. The integration of FBSMDs and our MIND dataset can be used for diet-level planning and healthcare that considers all information on nutrition, ingredients, their compounds and flavors, menus, and diets. Molecular consideration is essential for healthcare purposes as exemplified in existing studies on anti-cancer food identification [78, 71]. The information on anti-cancer foods can be integrated with our MIND dataset to design diet recommendations for cancer patients. Recent studies have identified the extent of the gut microbiome's importance to health and for precision diet development [10, 77]. We described our related clinical study in Task 4 "Clinical studies with the MIND dataset". Foods are critical to the prevention, management, and treatment of diseases. Our work will contribute by extending existing food science to diet-level planning and healthcare with machine learning to include recommending scientifically healthy and contextually attractive menus to patients resulting from consideration of the molecular level to the diet level.

ORxML framework for the ML application to support professional tasks Despite its dramatic success in many fields from academia to industry, ML still has critical shortcomings. In particular, a huge amount of data needs to be used when training the ML model so that the model can subsequently attain an acceptable level of performance. Given this necessity of creating high-quality datasets for ML research and development (R and D), government bodies and large companies have invested extensive financial resources for dataset construction projects involving ML. In 2021, the South Korean government allocated 260 million USD to construct high-quality datasets for transportation, healthcare, agriculture, manufacturing, and other domains in which the types of data include sensor values, text, sound, images, and video. These datasets will be released to the public via the AI Hub ($https : //aihub.or.kr/$). Subsequently, R and D projects will explore AI as a means of assisting specific human tasks such as driving, diagnosis, sports, and machine control. Our diet-level data synthesis case demonstrated in this study has been motivated by the necessity of assisting the diet design tasks of dietitians and physicians.

Meanwhile, unlike the other data construction projects that require manual effort, our case is unique in its use of a systematic approach integrating the different capabilities of input from experts with OR and ML. To avoid training ML models with readily available but poor-quality records of human tasks related to diet planning and DHR, we devised the ORxML framework to generate high-quality synthetic data for such tasks (Section 3). This framework can be used for other cases requiring the generation of high-quality synthetic data. In particular, the framework can be generally applied to address the data insufficiency problem of many “professional tasks” that are difficult to perform or describe, even among experts. The quality of some professional tasks, such as diet planning and drug discovery, cannot be easily trusted because even experts, including scientists and physicians, are not skilled enough to optimally execute these tasks. This challenge is attributed to the inherent complexity of a task that involves both explicit and implicit criteria. Given this challenge, ML models trained on available but poor-quality records of the task cannot be fully utilized because the model will not perform well. In addition, the collection of data on specific professional tasks, such as the acquisition of disease diagnostic records of medical doctors and product inspection records of manufacturing engineers, is not always economically viable. Moreover, the development of a simulation environment to generate realistic data of these tasks can be nearly impossible because of task complexity. The value of learning and automating professional tasks is high, given the difficulty and importance of these tasks, and requires assistance from machines. As described in Section 3, our proposed ORxML framework addresses this problem by integrating three modules with a “human-in-the-loop”: 1) the OR module to generate initial data satisfying the explicit criteria of a professional diet planning task; 2) professional dietitians who evaluate and adjust the initial data considering the implicit criteria of the task, i.e., criteria that cannot be modeled; and 3) the ML module to control the diet generation and refine the quality and completeness of the data.


Motivations of this work
Original contributions of this work
Beneficiaries, Values, and Implications of this work
Dietary healthcare research (DHR)
Food Science
Nutrition Science
Medical Science
Sport Science
Military Science
Menu & Ingredient
Nutrition
Disease
Performance
Injury
Integrated DHR DB
Diet
Diet planning with ML (diet sequence level)
composition of diet
individual preferences of diet
cultural & personal contexts of diet
Diet planning (menu item level)
combination of menus
choice of substitutes & complements
Diet problem (food item level)
the nutrition of items
per-food quantity
the cost of items
Diet planning and dietary healthcare research with ML
X
Poor quality data
X
Difficulty of diet planning tasks even to the professionals
K-MIND (Korean Menu-Ingredient-Nutrition-Diet dataset)
Diet (generated)
Menu (edited)
Diet data
Menu data
Ingredient (collected)
Nutrition (selected)
Ingredient data
Nutrition Guide
Ministry of Food and Drug Safety
Rural Development Administration
어린이급식관리지원센터
Center for Children's Foodservice Management
OR Module
Expert Module
ML Module
1) Formulate
2) Filter
3) Edit
4) Refine
5) Generate
6) Explain
Domain knowledge
Low-quality diet dataset
High-quality diet dataset
Decoder
"Start"
"End"
Machine learning
High quality data
Difficult professional tasks of humans

① Despite recent attempts on developing ML applications for diet planning and dietary healthcare research, there is no high-quality diet dataset, due to the complexity and difficulty of this job even to the professionals such as dietitians and physicians

② It is challenging to create a high-quality synthetic data of professional diet planning tasks

① We create and publish a diet dataset for diet planning and healthcare research with ML, called K-MIND, which include the information of Korean diets, menus, ingredients, and nutrients

② In creating the K-MIND dataset, we devised the framework for generating high-quality synthetic data of professional tasks, which can be used in other studies

① Our published K-MIND dataset will contribute to the dietitians and physicians concerning efficient and precise diet planning for patients and people, as well as to promote the use of ML for dietary healthcare research

② Our ORxML framework used for the data creation will contribute to the ML researchers who deal with real-world applications of ML for professional tasks

Figure 6: Summary and contributions of this work

Summary of this work on the MIND dataset and the ORxML framework In conclusion, the summary and contributions of this work are shown in Figure 6. (1) Despite recent attempts to develop ML applications for diet planning and DHR, there is no high-quality diet dataset due to the complexity and difficulty of this endeavor even for professionals. (2) The creation of high-quality synthetic data of professional diet planning tasks is challenging. (1) Thus, we have developed and published a diet dataset for diet planning and healthcare research with ML, called MIND, which includes information on Korean diets, menus, ingredients, and nutrients. (2) In creating the MIND dataset, we devised the ORxML framework for generating high-quality synthetic data regarding professional tasks, a method that can be used in other fields as well. (1) Our published MIND dataset will contribute to the work of dietitians and physicians concerning efficient and precise diet planning for patients and individuals and will promote the use of ML for DHR. (2) Finally, our ORxML framework used for data creation will contribute to ML researchers working with professional task applications of ML.

The contribution point (1) is significant and meaningful because diet planning has been mainly considered an academic operations research (OR) problem, particularly a combinatorial optimization problem, one that mainly aims to fulfill the nutrient requirements of diets. However, this approach has been ineffective in actual practice due to its limited capability to consider the implicit patterns in diets. These implicit patterns are considered most important by professional dietitians (see Section 1 and Appendix A.4). In our paper, we originally clarified that the diet planning problem should actually be considered a sequence generation problem in order to accommodate the implicit patterns in diet sequences, i.e., to fulfill the "compositional" requirements of diet planning and to effectively address the issue of having no high-quality dataset to promote "diet planning with ML."

The history of diet planning research is quite long. Around 1945, diet planning was first defined as a linear programming problem to identify optimal quantities of food items. However, humans do not consume a specific quantity of each food unit, a combination of cooked ingredients, but rather the end-product, the food "menu," as a whole unit. Thus, around 1990, the problem was expanded to a mixed-integer programming problem to identify optimal combinations of menu items. All previous studies considered ingredient- and menu-level information, but diet-level planning should involve the compositional patterns of menus in diets. The compositional patterns are implicit depending upon the contexts and should be addressed by data-driven machine learning (ML) approaches. This work appears to be the only study on diet-level planning, and the main reason for this is the lack of a high-quality diet dataset with which to investigate data-driven ML approaches. To the best of our knowledge, existing studies on the application of ML to dietary healthcare only considered the ingredient and menu levels.

In summary, diet planning is an important problem that should be solved with ML but could not be addressed in this way due to the lack of datasets for this data-driven approach. Our dataset is the first high-quality resource to promote the research field of "diet planning with

ML." Through research and development of the use cases exemplified in Appendix A.7, our goal is that the proposed dataset will be widely disseminated for diet planning and dietary healthcare with ML.

Finally, our dataset creation framework (ORxML) is unique and relevant to the dataset creation efforts for ML research. As such, our contribution point (2) is to devise the ORxML framework to systematically create a large-scale high-quality synthetic diet dataset. Given the high complexity of diet planning (see Section 1 and Appendix A.4), practical, high-quality diet data are lacking even though previous data were generated by professional dietitians. Thus, we created a novel high-quality dataset and devised an OR–Xperts–ML (ORxML) framework for diet planning and dietary healthcare with ML. This framework integrates the capabilities of OR, Expert, and ML modules. The OR module, a combinatorial optimization model, generates synthetic diets to satisfy explicit nutrient requirements. The Expert module evaluates and adjusts the initial data in terms of implicit composition requirements, the criteria that cannot be specified in the combinatorial optimization model; and the ML module automatically augments the data that ensure composition compliance with nutrition enhancement. A series of experiments demonstrate the significance of the three modules and validate the quality of our dataset (see Section 4 and Appendix A.6). Note that this framework can be used in any other contexts of creating diet data and those of difficult professional tasks as well.

A.8 Datasheets for Datasets

The following questions and answers are from the datasheets for the datasets framework [23].

1. **Motivation**

   (a) **For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.**

   A diet is important to all people from children to seniors and from healthy people to patients. As described in the introduction and literature review sections, the necessity of creating a benchmark dataset for DHR with ML has increased rapidly for food technology, nutrition management, clinical medicine, sports science, and military nutrition, among others. In addition, another important motivation of our work was as follows: In South Korea, most daycare centers rely on the local government's Center for Children's Food Service Management for diet planning. However, the dietitians employed in the government centers or daycare centers are burdened with designing diet plans because of the complexity of diet design. In addition, they have other duties, such as monitoring the cooking and hygiene status, as well as budget management. Our work was initiated to solve this problem and help dietitians efficiently design high-quality diets for children.

   As such, under the support from the Korean government, we aimed to develop AI that could automatically generate diets for diet planning and DHR in South Korea. However, to train the AI, we needed information on the ingredients of the menu and the nutrition corresponding to those ingredients. Moreover, the sources that publish the food and ingredient data are different, leading to inconsistencies in the data. As a result, it was difficult to have sufficiently organized data for training. Thus, datasets had to be developed before developing the diet-related AI, and so we worked with dietitians from hospitals to unify the ingredient data and menu data from different sources. We believe that generating diets based on the unified data that has been validated by dietitians will be a valuable asset for AI developers working on diet-related projects, in addition to our own research. Therefore, we decided to generate and distribute this data.

   (b) **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

   Authors from the Ulsan National Institute of Science and Technology created prototypes of the unified datasets using an operations research model for diet planning. Authors from the Kosin University College of Medicine validated and modified the prototype data from the nutritional and clinical perspectives through a collaboration with external experts. Then, based on the diet data refined by the experts, we trained a controllable generation machine to generate diet data. Further details on the generation process are shown in Section 3 in the main body and Appendix A.2.

(c) **Who funded the creation of the dataset? If there is an associated grant, please provide the grant name and number.**
This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No.2020-0-02135: Development of a gut microbiome-personalized diet recommendation AI system for the children with atopic diseases, No.2020-0-01336: Artificial Intelligence Graduate School Program - UNIST). This work was supported by the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Ministry of Science and ICT (grant number 2019M3E5D1A02070867) and by the Ministry of Education (grant number NRF-2021R1I1A4A01049121). This work was supported by the 2021 Research Fund (1.210050) of UNIST.

2. **Composition**

(a) **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)?**
Our dataset includes instances of food ingredients, menus, and diet. See the Supplementary Material for more detailed information.

(b) **How many instances are there in total (of each type, if appropriate)?**
Our dataset includes 3,036 ingredient instances, 3,238 menu instances, and 1,500 daily diet instances.

(c) **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**
Yes. Our dataset contains all possible instances.

(d) **What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.**
The ingredient data involve the category and nutrition information for each ingredient item. The menu data involve the category, note, and ingredient information for each menu item. The diet data involve the identifier and its menu composition. See the Supplementary material for further details.

(e) **Is there a label or target associated with each instance? If so, please provide a description.**
Yes. For example, menus are labeled with their category (e.g., rice, soup). See the Supplementary material for further details.

(f) **Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.**
Yes. From the original source of the government database, various types of nutrients have been removed due to the incompletion issue (e.g., Amount unknown for specific ingredients). We included the most prominent types of nutrients based on the literature of nutrition.

(g) **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.**
Yes. For example, menus are related in diets, while ingredients are related in menus. See the Supplementary material for further details.

(h) **Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.**

Although we tried hard to minimize any errors, sources of noise, or redundancies in the dataset, there could be such cases. Meanwhile, two or more recipes may exist for one kind of menu. Because there is a nutritionally significant difference between the two recipes, we did not select one recipe only. We included both cases by differentiating the minor part of the menu name (e.g. From 'Pork fried rice made with oyster sauce' to 'Pork tenderloin fried rice made with oyster sauce').

(i) **Is the dataset self-contained, or does it link to or otherwise relyon external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources.**
The MIND dataset is self-contained.

(j) **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.**
No.

(k) **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.**
No.

3. **Collection Process**

(a) **How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)?**
The ingredient data had been collected and created from the Rural Development Administration of the Korean government through validated experiments. The menu data was developed by the Children's food service management center under the Ministry of Food and Drug Safety of the Korean government based on the expertise of the affiliated professional dietitians. The diet data were generated in this work. See the main body of this article for details.

(b) **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?**
The ingredient and menu data could be downloaded from the Web pages of the government organizations. As aforementioned, we created high-quality synthetic diet data based on the proposed method in this work.

(c) **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**
The ingredient and menu data were collected by the authors. They created the diet data.

(d) **Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.**
The data collection and creation jobs have been conducted from June 2020 to August 2021.

(e) **Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.**
The corresponding authors (one affiliated with an engineering school and the other with a medical school) confirm that our data do not need an ethical review

process. The aforementioned grants do not require an ethical review process regarding our data.

(f) **Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.**
No.

4. **Preprocessing/cleaning/labeling**

(a) **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.**
We conducted preprocessing and cleaning jobs for the ingredient and menu data sourced from the government organizations. See the Supplementary material for further details.

(b) **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.**
No.

(c) **Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.**
No.

5. **Uses**

(a) **Has the dataset been used for any tasks already? If so, please provide a description.**
Yes, our work has already started to create a real impact. As the synthesized high-quality diet-level dataset MIND is intended to be used in the development of a machine for professional dietitians and pediatricians, the Center for Children's Food Service Management in South Korea have indicated that they will use our outcomes from the fall of 2021. In addition, a startup company in South Korea will use this dataset for the development of a gut microbiome-personalized diet recommendation AI system for children with atopic diseases and food allergies. See Appendix A.7 for further details.

(b) **Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.**
No. We published our MIND dataset in August 2021 for the first time.

(c) **What (other) tasks could the dataset be used for?**
Our data and its management tools are useful for many food-related tasks that require the verification of ingredients and nutrients. For example, data can be used to filter menus containing allergen ingredients from a diet for allergic patients. In addition, a diet generation application to support diet planning for allergic patients can be trained based on our MIND dataset. See the Introduction and Appendix A.7 for further details.

(d) **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?**
The diets provided by MIND are healthy "reference diets" for an unspecified majority of the population. Therefore, there is no guarantee that a user will have a positive response to the provided diets for their meals. Although our diets have been evaluated and confirmed by the related government organization and the affiliated professional dietitians, there is a further need to consider health and preference issues, such as food allergies and low-salt diet preference, when serving

the provided diets. We do not consider these user-specific factors because our dataset was not created for a particular target group for a customized purpose. Therefore, in order to use the diets provided in this MIND, it is recommended to inspect the ingredients and nutrition using the analysis functions provided together in the dietkit package.

(e) **Are there tasks for which the dataset should not be used? If so, please provide a description.**
We cannot imagine such tasks yet. There should be no problem, given that the original source data from the government are all certified, while the synthesized data were created by machines and confirmed by professional dietitians.

6. **Distribution**

(a) **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.**
No. We distribute the data to the public. See the last part (license and rights) of the Supplementary material.

(b) **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?**
Our dataset and its management package have DOI: 10.5281/zenodo.5302044
Also, our dataset is distributed with the management package though GitHub: github.com/pki663/dietkit and PyPI(The Python Package Index): pypi.org/project/dietkit

(c) **When will the dataset be distributed?**
It is currently available.

(d) **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.**
The MIND dataset has three sub-datasets with different license information. See the last part (license and rights) of the Supplementary material.

(e) **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**
No.

(f) **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**
Regulatory restrictions can be applied to the data depending on the case and the type of instance. See the last part (license and rights) of the Supplementary material.

7. **Maintenance**

(a) **Who is supporting/hosting/maintaining the dataset?**
The MIND dataset is hosted by the two channels: GitHub and PyPI (The Python Package Index). See the Supplementary material for the data maintenance plan.

(b) **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**
By email: sooo@unist.ac.kr, chlim@unist.ac.kr, or my.jung@kosin.ac.kr

(c) **Is there an erratum? If so, please provide a link or other access point.**
No.

(d) **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?**
Yes. We will update the data approximately every year as relevant data sources are updated. For example, the National Standard Food Components, which is the original source of the ingredient data, is updated every year by the Rural Development Administration of the Korean government.

(e) **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances.**
The MIND dataset does not relate to people. Therefore, there is no limit on the retention of the data associated with the instances.

(f) **Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.**
Yes. Older versions are automatically preserved and provided by the GitHub host. Users can access previous versions at any time.

(g) **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.**
First, see the last part (license and rights) of the Supplementary material for the potential extension of our dataset. Second, we will accommodate the suggestions and contributions for our dataset from the users through GitHub, and all amendments will be explicitly recorded as commit commands. Users can upload or connect their data to our dataset through GitHub. Meanwhile, we would like to maintain and control the structure of our dataset (although we will accommodate suggestions), such that the data and contents are organized and managed consistently and coherently.