Advanced Engineering Informatics 51 (2022) 101516

Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei

Full length article

Counterfactual inference to predict causal knowledge graph for relational transfer learning by assimilating expert knowledge --Relational feature transfer learning algorithm

Jiarui Li [a,*], Yukio Horiguchi [b], Tetsuo Sawaragi [a]

[a] Department of Mechanical Engineering and Science, Graduate School of Engineering, Kyoto University, Kyoto, Japan
[b] Faculty of Informatics, Kansai University, Osaka, Japan

ARTICLE INFO

Keywords:
Transfer learning
Causality
Counterfactual inference
Knowledge networks
Explainable machine learning

ABSTRACT

Transfer learning (TL) is a machine learning (ML) method in which knowledge is transferred from the existing models of related problems to the model for solving the problem at hand. Relational TL enables the ML models to transfer the relationship networks from one domain to another. However, it has two critical issues. One is determining the proper way of extracting and expressing relationships among data features in the source domain such that the relationships can be transferred to the target domain. The other is how to do the transfer procedure. Knowledge graphs (KGs) are knowledge bases that use data and logic to graph-structured information; they are helpful tools for dealing with the first issue. The proposed relational feature transfer learning algorithm (RF-TL) embodies an extended structural equation modelling (SEM) as a method for constructing KGs. Additionally, in fields such as medicine, economics, and law related to people's lives and property safety and security, the knowledge of domain experts is a gold standard. This paper introduces the causal analysis and counterfactual inference in the TL domain that directs the transfer procedure. Different from traditional feature-based TL algorithms like transfer component analysis (TCA) and CORelation Alignment (CORAL), RF-TL not only considers relations between feature items but also utilizes causality knowledge, enabling it to perform well in practical cases. The algorithm was tested on two different healthcare-related datasets — sleep apnea questionnaire study data and COVID-19 case data on ICU admission — and compared its performance with TCA and CORAL. The experimental results show that RF-TL can generate better transferred models that give more accurate predictions with fewer input features.

1. Introduction

Transfer learning (TL) is a machine learning (ML) method in which knowledge is transferred from the existing models of related problems to models for solving the problem at hand. In real-world applications, such as in the field of healthcare, sometimes large numbers of labeled instances are difficult to collect. Solving the problem of limited labeled data is one of the applications of TL. TL involves using relationships between features (X_S) in the source domain (D_S) and features (X_T) in the target domain (D_T) and transferring the model from the source task (T_S) to the target task (T_T) [1]. TL can be classified into four categories: instance-based, parameter-based, feature-based, and relational. In this study, we focus on the latter two types.

Feature-based TL methods consist of finding the statistical characteristics of data distributions and projecting data features in D_S and D_T into a particular data space by making a feature transformation such that the difference between the statistical characteristics expressed in D_S and D_T is reduced [2]. The key to the methods in this category is to choose a suitable statistical characteristic of the data distribution. For example, the classic feature-based TL, Transfer Component Analysis (TCA),

Abbreviations: AHI, Apnea Hypopnea Index; CEA, Assumption of Constant Effect; CFA, Confirmatory Factor Analysis; CORAL, CORelation Alignment; EFA, Exploratory Factor Analysis; EMR, Electronic Medical Records; FPCI, Fundamental Problem of Causal Inference; GA, Genetic Algorithm; GoF, Goodness of Fit; HA, Assumption of Homogeneity; ICU, Intensive Care Unit; KG(s), Knowledge Graph(s); LTL, Language-bias Transfer Learning; ML, Machine Learning; OSA, Obstructive Sleep Apnea; RF-TL, Relational Feature Transfer Learning; SEM-EML, Structural Equation Modeling-Explainable Machine Learning model; SF_36, 36-Item Short Form Survey; SUTVA, Stable Unit Treatment Value Assumption; TCA, Transfer Component Analysis; TL, Transfer Learning.

* Corresponding author.
E-mail address: ljr10225008@gmail.com (J. Li).

https://doi.org/10.1016/j.aei.2021.101516
Received 24 April 2021; Received in revised form 18 November 2021; Accepted 24 December 2021
Available online 31 December 2021
1474-0346/© 2022 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

proposed by Pan et al. [3] transforms data features into a Hilbert space where the distance of the data features' marginal probability distribution can be minimized. Sun et al. [4] developed CORelation Alignment (CORAL), which transforms data features from D_S to D_T by minimizing the difference between the covariances in each domain. Similar to [3,4], a variety of criteria have been explored for feature-based TL [5,6,7]. However, these methods share certain shortcomings. First, whether it is transforming data features of D_S and D_T to the same space or transforming data from one domain to another, D_S and D_T's data features can be similar but not identical. That is, the transformation process more or less loses the original feature attributes, which may result in the trained target-domain model having poor accuracy. Compared with feature transformation, it would be better to select data features in the target domain according to the source domain's information and the relationship between D_S and D_T.

Additionally, the methods mentioned above only consider the relations of data features between the source and target domains; they ignore relations among the same domain's data features. There is usually little difference in the statistical distributions of data features between the source and target domain in practical applications. When large numbers of data features exist in both the source and target domains, the above methods often transfer all data features in D_S to D_T and fail to reduce X_T's dimensionality because of the difficulty in telling the difference between the data distributions of D_S and D_T. However, in machine learning, a simple prediction model with fewer data features is always preferred to a complex one with many features. Although methods such as TCA allow users to define the number of features to be transferred manually, this is hard to do without prior knowledge. Besides relations across domains, knowledge of the dependences among data features in the same domain is also essential. An effective machine learning model requires the data features used for prediction to be closely correlated with each other and the target of prediction. Data features that do not meet this requirement are unnecessary for the target domain model, even though they may have a similar distribution with the source domain's corresponding data features. It is critical to clarify the relational structure of the source domain's data features before transferring them to the target domain. When the relations in a model are the objects to be transferred, the methods are called relational TL.

Relational TL accounts for the relationships among data features, and the transferred objects are the logic networks in D_S. It assumes that the knowledge networks in D_S and D_T are the same or can be transferred from D_S to D_T. Two critical issues affect the development of relational TL: 1) how to extract knowledge networks from the data of the original domain and how to transfer knowledge networks from one domain to another.

Knowledge graphs (KGs) are useful tools for dealing with the first issue. KGs are knowledge bases that use data and logic to structure information. They are often used to store interlinked descriptions of entities with free-form semantics [8]. KGs express not only statistical relationships among data but emphasize the human reasoning involved in knowledge representation. According to [9,10], knowledge-based modeling manipulations are categorized into ontologies, cognitive knowledge bases, linguistic knowledge bases, and expert knowledge bases. Although expert-knowledge-based modeling methods have been criticized for their heavy reliance on expert experience, such experience and knowledge constitute an indispensable gold standard for validating models [11,12,13]. Peng and his team [14] proposed a hyper-network-based approach to retrieve data and reasoning with engineering design knowledge. Bayesian inference has been used for constructing the KGs. In [15], a Bayesian network with noisy OR gates was used to extract a health knowledge graph from Electronic Medical Records (EMRs). Bayesian-based technologies have been widely utilized for making KGs because of their intuitiveness and interpretability. However, Bayesian-based models depend on probabilistic inference, which cannot explain the correlations and causalities among data; this limits their application to KGs involving causal logic.

Structural equation modeling (SEM) is a well-known data modeling method expressed by a series of regression functions. The original SEM cannot directly be used as an inference model. Our team proposed an explainable ML model based on SEM (SEM-EML) [16]. In the present study, we referred to the key procedures in SEM-EML and introduced SEM in TL technologies to extract KGs from data as a preparation for transfer learning.

Another essential issue with relational TL is the ways of transferring. Unlike instance-based, parameter-based, and feature-based TL, the difference between D_S and D_T is easily expressed mathematically, such as the distance between data features across domains. However, the difference in the relational structure between D_S and D_T is hard to describe statistically. That is, the transference of relation needs support from a human expert. To the best of our knowledge, there are few algorithms for transferring "relations" [17,18,19]. Kumaraswamy et al. [19] developed an interactive TL algorithm in relational domains, called language-bias transfer learning (LTL), that uses tree-type inductive logic programming. The transference procedure of LTL entirely depends on a human expert's experience assisting the algorithm in selecting appropriate relations to transfer, which is time-consuming and laborious. Instead of interacting with expert experience, a more efficient way is to teach algorithms to imitate human cognition. A number of cognitive factors have been identified as being involved in the support of transferring empirical engineering knowledge [20]. In particular, causality, as a human inference logic, has attracted attention from researchers as ways for assisting and directing machine learning. Analogical reasoning, the well-known feature-mapping method proposed by Gentner and his team [21], is a helpful tool for inferring relational structures from one domain to another. Gentner discussed that attention to the differences in objects between domains leads to the inference on the relationships among the objects. Through the procedure of analogy, features in one domain can be mapped to another one. Gentner's method stresses the similarities of relational structures in different domains. However, distinctions between domains were ignored. The mapping or transference should not be a static contrast but rather a dynamic process. In the presented study, we take advantage from another aspect of causality, counterfactual inference, which is able to guide the dynamic process of feature transference across domains.

A causal relationship is recognized as ground truth, and a change in the reason will cause a corresponding change in the result. In machine learning, the reason is a stimulus given to a model. The result is a change in the model produced by the stimulus. Furthermore, in causality theory, a prediction that if the same stimulus is experienced in the future, the model will change is called a counterfactual inference. The task of relational TL algorithms is to predict the unlabeled target in a domain by transferring a relational structure from another domain. If the relational structure changing rules from the source domain to the target domain can be inferred from a piece of particular causal knowledge, it will be feasible to predict a model in the target domain. Using causality as a guide for the learning procedure in ML is efficient because the only information supplied by a domain expert is a piece of causal knowledge. The causal TL algorithm proposed by Rojas-Carulla et al. [22] uses SEM for finding the invariant domain between D_S and D_T. Roughly speaking, the algorithm uses the invariance of the reasons in a causal relationship to find conjunct causal features in the two domains. However, it focuses on how to extract causalities, not how to transfer knowledge.

This paper proposes to use counterfactual inference to predict causal knowledge graphs from the source domain to the target domain for relational transfer learning. We name the algorithm we use for inference Relational Feature Transfer Learning (RF-TL). The counterfactual inference is made according to the causal knowledge provided by a domain expert, which predicts the relations among X_T from the relations in X_S. Moreover, other ML methods are used to label the data in D_T using the extracted features.

The contributions of this paper are as follows: 1) Different from the traditional feature-based TL algorithms that are limited to differences in

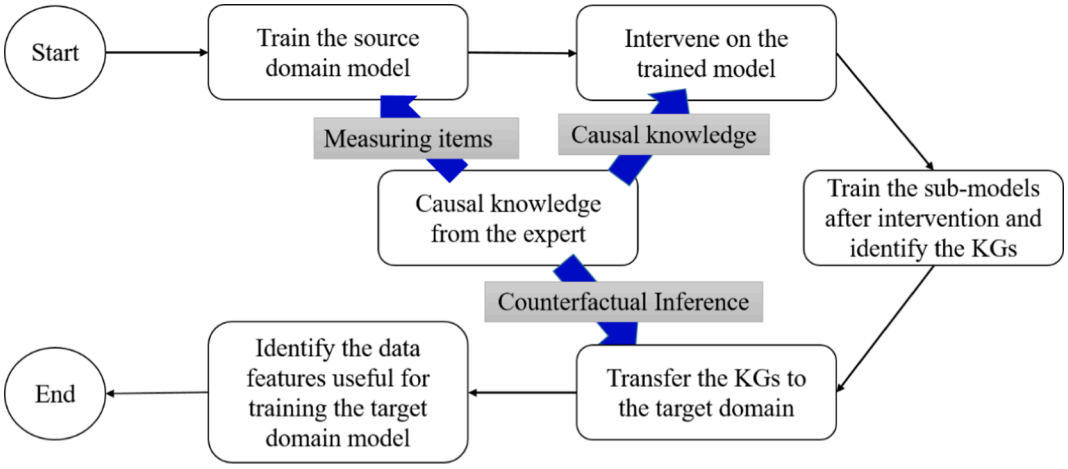


Fig. 1. Overall design of RF-TL.

data features across the domains, RF-TL takes into account dependency relationships among data features in each domain, thereby facilitating the extraction of functional features and eliminating useless ones. 2) RF-TL does not require feature transformation. Instead, it selects appropriate features in the target domain to train a model. This preserves the feature attributes as much as possible. The experimental results show that RF-TL can use fewer data features to train a model that is more accurate than other models made with the traditional feature-based TL methods, TCA and CORAL. 3) The introduction of human expert knowledge makes the transference explainable and reasonable. Our pioneering utilization of counterfactual inference in TL makes the learning process accurate and efficient.

The remainder of this paper is organized as follows. Section 2 gives technical background, including an overview of relations between KGs and SEM, and explains how SEM can contribute to constructing KGs for domain knowledge. It also gives a brief introduction to TCA and CORAL, which we selected as methods for comparison in the experiments. Section 3 describes the procedure of RF-TL. Section 4 describes an experiment we conducted to evaluate the effectiveness of applying RF-TL to healthcare cases, and Section 5 discusses the results. We give concluding remarks in Section 6.

2. Technical background

2.1. Structural equation modeling for constructing knowledge graphs

The domain knowledge that is used for solving problems is expressed as rules in KGs. The rules are made up of IF and THEN parts. The IF part can include first-order logic expressions, e.g., the conjunction AND or disjunction OR. Nodes in KGs consist of linguistic objects and their values. Rules represent relations among nodes and can be classified as logical or fuzzy [23]. At present, domain knowledge is mostly acquired from domain experts, while automatic or semi-automatic methods have been proposed for saving labor and time [24,25].

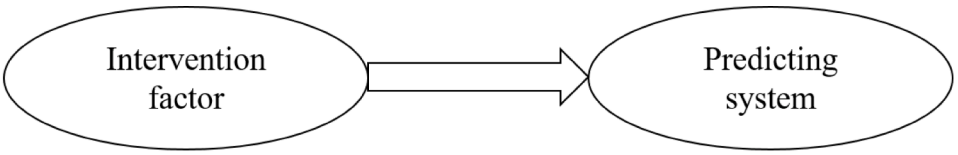
In our previous study, we proposed an SEM-based machine learning method named SEM-EML for inferring engineering problems [16]. The procedure of SEM-EML is described in the Appendix. In the present study, we use SEM-EML to obtain the structure of KGs from empirical data. As a way of measuring correlations among data points, the utilization of SEM for constructing KGs makes it possible to add properties to "edges", i.e., IF A is *a* AND B is *b*, THEN A strongly (weakly) results in B, which extends the usable range of KG expressions on domain knowledge. Also, SEM's strong point over other information integration methods, e.g., Bayesian networks, is its ability to measure causalities between factors (corresponding to nodes in KGs). The notion of causality lets a static binary relationship between nodes, e.g., IF A is True, THEN B is true, acquire dynamic properties, e.g., IF A changes, THEN B will change. Dynamic properties are essential to KGs, without which KGs can only store and express "data from the past" but never predict the future. As the application in this study, we describe transference as a dynamic procedure that requires KGs to cope with change.

2.2. Feature-based transfer learning methods

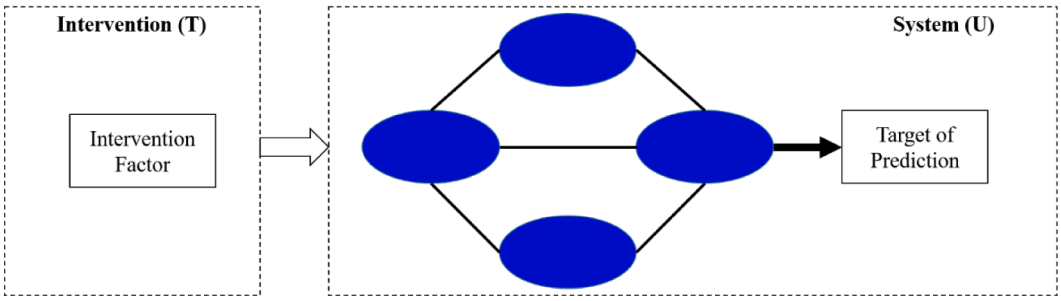
Besides knowledge network extraction, another critical problem of transfer learning is how to transfer the relationships from the source domain to the target domain. As mentioned in Section 1, several methods can be chosen depending on the transfer objects. In this study, we focus on feature-based transference. TCA and CORAL are representative feature-based TL methods and are briefly introduced here. Section 4 describes experiments that compared their performance with that of the proposed algorithm.

TCA maps data features in D_S and D_T into a high-dimensional reproducing kernel Hilbert space, where the distance between the data features in the marginal probability distributions over D_S and D_T is minimized while preserving their respective internal properties to the greatest extent. TCA extends the principal component analysis to TL, and TCA and PCA's core ideas are similar. In the transformed feature space, only the principal components are needed to be preserved. We call this idea dimensionality reduction. As mentioned in Section 1, although the user can decide the number of dimensionalities that remains after TCA, it is hard to choose an appropriate number without prior knowledge. Also, the number of dimensions influences the accuracy of learning to a great extent. Our experiment in Section 4 shows how the decision on the dimensionality number affects learning accuracy. Moreover, we show that RF-TL does not have this selection problem.

Different from TCA that transforms data features in both D_S and D_T into another space, CORAL transforms only X_S to D_T and uses the transformed X_S to train a model in D_T. The basis of CORAL is to extract correlations among data features and then transform the covariance matrix from the source domain to the target domain. On the one hand, while the distributions of data features are not so different from one domain to the other and they correlate strongly in each domain, CORAL fails to reduce the dimensionalities. On the other, two data features with strong correlations do show they have a particular relationship with each other while no causal relationships are interpreted. Correlations cannot tell us how one data feature changes in correspondence to a change in another data feature's change. While it is not a problem to use data features with solid correlations to train a machine learning model, in TL, the transference is an automatic procedure. It is necessary in TL to predict the change in a model when data features in the source domain are changed to the target domain.



(a). Concept of interventionism-causality



(b). Mechanism of interventionism-causality

Fig. 2. Intervention-causality theory.

Thus, the learning structures expressing causal knowledge must be known in the source domain so that the correct transference of the model to the target domain can be conducted. Here, SEM is an excellent tool for extracting causal knowledge from data, and it is used in RF-TL.

3. Relational feature transfer learning algorithm (RF-TL)

The overall design of RF-TL is shown in Fig. 1. The core idea of RF-TL is to use causality to direct the counterfactual inference from D_S to D_T. An explainable model structure is necessary regardless of whether one is conducting the causal analysis or counterfactual inference. For training the source model, expert knowledge should be used as a measuring item (s) of the model so that in the next step, an intervention can be performed on the model. Furthermore, after extracting the knowledge network from the intervened sub-models, RF-TL uses counterfactual inference to predict the KG(s) carrying the information on features useful for D_T. The next sections illustrate the specific procedures of each step, including the role of the causal relationships among them.

3.1. Interventionism-causality knowledge of domain experts

Causality is a philosophical concept. When two events occur in certain time order, one event has an impact on the other. The event occurring earlier is the reason and the event occurring later is the result. An "Order" is very important for actual causality [26]. Introducing causality theory to ML usually involves adopting the interpretation of interventionism. In interventionist-causality theory, an intervention is regarded as a cause, and the corresponding changes in the system are the results [27]. Fig. 2(a) shows the concept of interventionism-causality.

Causality is regarded as a factual truth in the real world. In a causal model, the direction of the arrow is non-reversible, which also clarifies the essential difference between causality and correlation. When we talk about two events being statistically correlated, we can only show that the two events have a particular relationship. However, there is no illustration about the "order" or which one impacts the other. In other words, causality is a ground truth or customary rule and is higher in some sense than the level of a statistical relation. In the interventionism-causality system, the intervening factor (the reason) is objectively variable and will lead to a corresponding change in the predicting system. There are many cases in real life where this theory applies. For example, the risk of getting a disease such as hypertension and diabetes becomes higher with increasing age. A change in a population will influence the economy. In a production line safety assessment system, the temperature of the environment is an essential factor affecting the safety risk.

However, a commonality of the above-mentioned cases is the bias in data collection caused by objective facts. Sometimes, the collection of global data is impossible or inhumane. For example, data on diseases that occur more frequently in older age groups are scarce from young people. It is impossible to artificially make the young age quickly to get an age-wide predictive system. Similarly, it is unrealistic to change the population structure of a society in a short time. However, using existing data and by taking advantage of interventionism, we can observe a change in a system caused by an intervention factor. Furthermore, we can transfer the model constructed using existing data to the domain in which we want to predict. The details of the interventionism-causality mechanism are shown in Fig. 2(b).

For the sake of illustration, suppose that we are to design an attendance forecasting system for baseball games. Baseball is usually not played in winter conditions, but the client wants to predict the attendance rate in winter. In this case, we define weather temperature as the intervening factor. Thus, the source domain D_S including data features in summer, and D_T represents the winter event.

Generally, in interventionism-causality, an intervention (T) is a stimulus applied to a system (U). The state (Y) of U changes in accordance with the stimulus. The intervention procedure is expressed as $\delta(u) = Y_t(u) - Y_c(u)$, where $Y_c(u)$ is the original state of U and $Y_t(u)$ is the state after the intervention. Using the baseball game prediction case mentioned above, we consider that temperature is the reason for the attendance rate. Then, if there is a system that can infer the attendance rate, the state of this system will respond accordingly to temperature intervention.

In practical applications, we would like to know the effect of an intervention on multiple systems, e.g., the effect of temperature on the decision to attend by a group of people. The following equation can be used to determine this effect

$$E[\delta(u)] = E[Y_t(u)] - E[Y_c(u)]$$

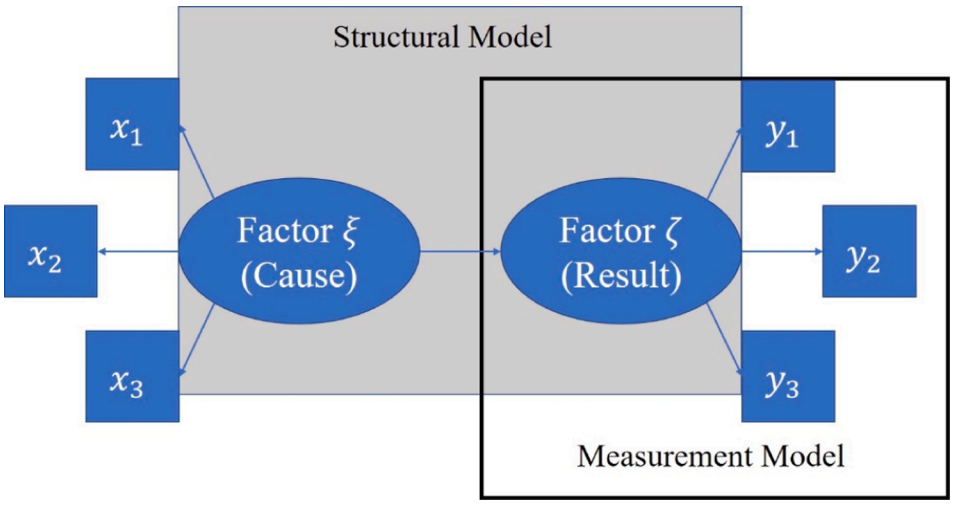


Fig. 3. Conceptual model of SEM.

where $E[\bullet]$ represents the average state of a group of individuals.

However, in practice, it is difficult to obtain accurate information on the state Y of a group of people, which is called the fundamental problem of causal inference (FPCI) [27]. In this case, it is impossible to ask every person in the world whether they would attend a game in winter. FPCI embodies the difficulty of determining $Y_t(u)$ and $Y_c(u)$ at the same time. In particular, three assumptions constrain the interventionism-causality [28]: A) the stable unit treatment value assumption (SUTVA) regards every individual change as an independent event; B) the assumption of constant effect (CEA) supposes that the effects of an intervention are the same for every individual. That is, $\delta(u_i) = \delta(u_j)$ if i and j are different individuals in the same group; C) the assumption of homogeneity (HA) is such that $Y_t(u_i) = Y_t(u_j)$ for two individuals. Under these three assumptions, it is easy to estimate the effect of an intervention on a group of objects. Our RF-TL follows these three assumptions.

The next step after constructing a causal model is to carry out counterfactual inference. "Counterfactual" means the fact has not occurred but can be predicted according to certain evidence. The most important message conveyed from the causal model is that a change in reason will cause a change in the result, but the reverse is not true. Therefore, counterfactual inference can be made as if the "reason" will change in the future, changing the "result" correspondingly. Coming back to intervention-causation, we could say that "if a certain intervention is carried out on a model, the system will obtain $\delta(u)$". Note that $\delta(u)$ only represents the change in the state, so it can be quantitative or qualitative. In the case of RF-TL, $\delta(u)$ is used as the transfer rule, which means it is qualitative. In the baseball game example, $\delta(u)$ can be obtained by intervening on temperature. Furthermore, counterfactual inference can be performed as "if there is an intervention on temperature, then the predicted attendance will change according to the rule (s)." Similar to the baseball game example, the main idea of RF-TL is to extract the "rule(s)" from the intervention conducted on the D_S model and make a counterfactual inference to transfer the knowledge network to D_T in accordance with the "rules".

In the following sections, we will describe the approach for KGs extractions using an SEM-based method. Then we will show the specific steps of RF-TL from training the source domain model to the transference of KGs from D_S to D_T.

3.2. Translating structural equation model into knowledge graph

As mentioned in Section 1, SEM is a valuable tool for digging into statistical causal relations in data. SEM is usually framed as a two-step procedure. The first step is exploratory factor analysis (EFA). The other is confirmatory factor analysis (CFA). EFA is a reliable tool for classifying data items into corresponding factors without a specific hypothesis, which aims to identify latent factors based on the observed variables. The measurement model and structural model make up the hypothesis for CFA to test. EFA yields extracted factors and their inclusive manifest variables that constitute the measurement model. The structural model specifies the logic paths among factors. Once the model is constructed, the factor loadings between manifest items and latent factors and between every two factors are estimated in accordance with the covariance matrix of the manifest items. Fig. 3 shows a conceptual graph of an SEM model.

In this study, we use SEM to construct KGs. Because the original SEM is a data analysis model, in order to use it to extract KGs, it has to be modified with several further operations.

First, we need to transfer SEM into a predictive system. The main steps are shown in the Appendix on SEM-EML. Roughly speaking, they include data collection, data management, structure management, and parameter learning. A common problem of SEM is that the validation of the model relies on a convincing hypothesis given by a domain expert, which is sometimes impossible or involves labor and time to obtain. In our approach, the strategy is adopted to optimize the structure of SEM. In the structure management procedure, to guarantee the model's validity, we use a genetic algorithm (GA) to identify the fittest model by setting goodness of fit (GoF) indexes. In the final step, the target of the prediction item is separated from other items by using a linear regression procedure.

Next, the obtained SEM-like predictive system is translated into KGs. The origin of using KGs can be traced back to the semantic network developed in the 1970s [29]. In particular, GOOGLE used a KG to enhance the performance of its search engine in 2012 [30]. There is no gold-standard definition for KGs, but they consist of a set of interconnected entities and their attributes [31]. In other words, a KG is made up of pieces of knowledge, and each piece can be represented as a subject-predicate-object relationship. The subjects and objects are the nodes in the graph, and a predicate is an edge describing the relationship between two nodes. The elements of the KG are defined as follows.

Definition 1. *Nodes: a) Body nodes are latent factors. b) An end node is the target item of the prediction, which also consists of a text description and label value.*

Definition 2. *Edges: a) Body edges are arrows connecting the body nodes and they represent the causal dependence between the nodes. An adjective word "Weak" or "Strong" is added to the edge as an attribute of the relationship. b) An end edge is an arrow pointing to an end node and it represents the predicate "predict", and it is not necessary to add the adjective pair.*

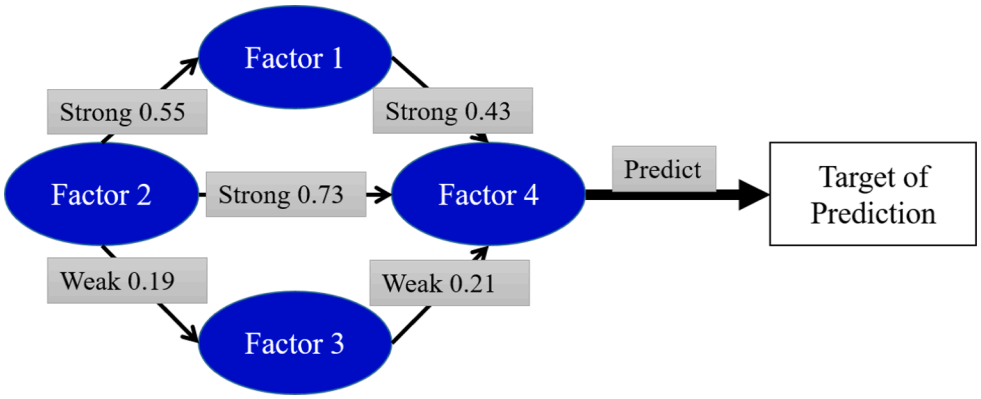


Fig. 4. SEM-like KGs. The model consists of latent factors, and measuring items belong to the factors (items except for the target of prediction are omitted in the figure). Each latent factor represents a node in the KG made up of an ontology expression and statistical values regressed from the items. Between the nodes, the arrows are the edges of KG with an ontology expression of Weak or Strong and a path loading value. The end node is the target of the prediction item, and an edge pointing to it expresses the action of prediction.

In Definition 2, the adjective word "Weak" or "Strong" is added to edges. The choice between "Weak" or "Strong" depends on the path loading (standardized path coefficient) between the nodes. "Weak" is given to edges that have path loadings (absolute value) < 0.3 between two nodes, while "Strong" is given to those with path loadings (absolute value) ≥ 0.3 (all the relationships should show statistical significance). In SEM, the path loadings evaluate the effect of one factor on the other. The factors that have a strong effect on each other are necessary for constructing the model. The path loadings are the standard regression coefficients between two nodes connected by an arrow, which relates to the (partial) correlation value. A model with a high goodness of fit means it can express the correlations among the factors comparably with the true relationships among the data, requiring the nodes connected by the arrows to have competing strong causal effects on each other. Although different researchers have different opinions on the reference point of the path loading [32,33], 0.3 is a safe choice. The effect of choosing different thresholds for the path loading is not a key point here. Users can choose a suitable number according to their application. The practical examples shown in this paper are medical cases, for which we chose 0.3 as a threshold for RF-TL to judge the "Weak" or "Strong" tags. If any factor has a low factor loading compared with all the other factors, it would be weak one in a prediction model. Fig. 4 shows the concept of a translated SEM-like KG.

As shown in Fig. 4, each ellipse represents a latent factor, and the items for measuring the factor are represented as rectangles. Note that, except for the target item of prediction, the other measuring items are not shown in the figure. The SEM-like KGs are made up of pieces of knowledge. For instance, in Fig. 4, Factor 3 is weakly related to Factor 4, and Factor 1 is strongly related to Factor 4.

Three predicate functions are used for expressing the knowledge in KGs:

$$S_i(x, y, fl_{x\to y}) \tag{2}$$

$$W_i(x, y, fl_{x\to y}) \tag{3}$$

$$N_i(x, y, 0) \tag{4}$$

Functions (2)–(4) represent three propositions. The subscript i in the functions represents the i^{th} sub-group, x and y are the nodes in the KG, $S_i(x, y, fl_{x\to y})$ means x results in y with a factor loading $fl_{x\to y,}$ and the relationship is strong, and $W_i(x, y, fl_{x\to y})$ means x results in y with a factor loading $fl_{x\to y,}$ and the relationship is weak. The order of x and y cannot be changed in Functions (3) and (4). Function (5) means there is no relation between x and y, where there is no arrow between the two nodes in the graph (the standard regression coefficient approaches zero).

3.3. Model training in the source domain

The first step is to train the predictive model for D_S. RF-TL only cares about strong/weak relationships between nodes of the intervened models. Thus, when training the source domain part, the knowledge expressing relationships on the edges does not have to be shown in the figure. In other words, only the procedures described in Section 3.2 that "transfer SEM to a predictive system" are conducted in the current step.

In this research, we only consider the situation in which the reason and result have a linear dependence. In the causal relation used by RF-TL, the "reason" is the intervention item. The "result" is the prediction target, and its target can be statistically expressed, such as the attendance rate of the baseball game.

In the source domain model, the item used as the intervening factor should be one of the measuring items of one of the latent factors, ensuring that the model and intervention are relevant. Once more using the baseball example, the temperature is the intervening factor, e.g., the "reason". A change in the intervening factor will cause a corresponding change in the prediction system P, i.e., the attendance prediction system. Then, the trained source model is constructed, as shown in Fig. 5.

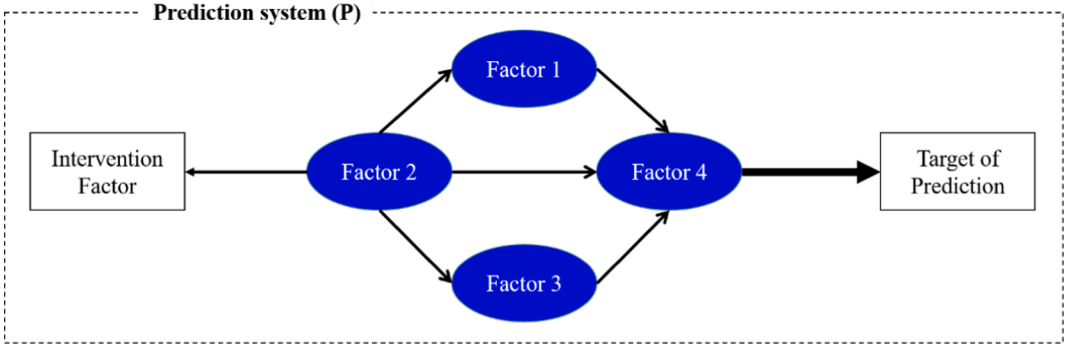


Fig. 5. The prediction system in the source domain. The intervening factor is a measuring item of factor 2, which is used as the intervention item, e.g., the "reason".

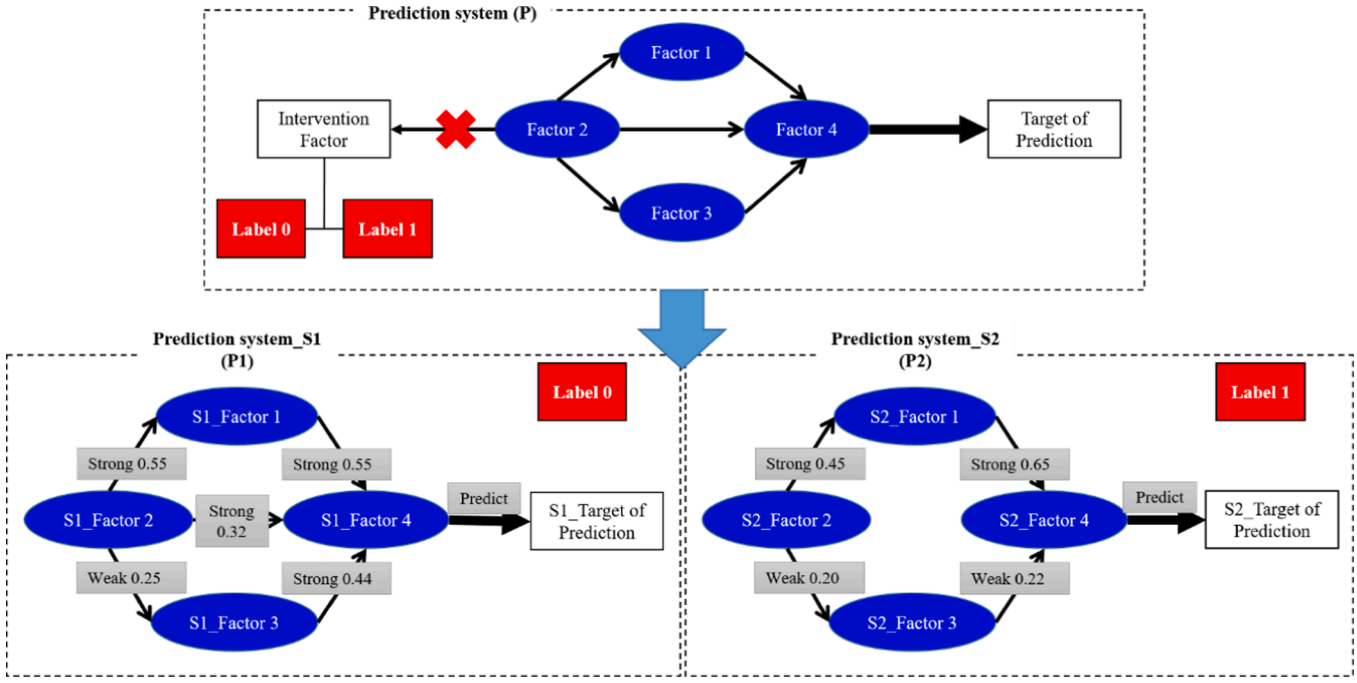


Fig. 6. Case of an intervention performed on the intervening factor and two trained sub-models. In the upper part of the figure, the red cross represents that the intervening factor is constrained to be constant label values, e.g., 0 and 1 in the example. After the intervention, the intervening factor is removed from the figure, and the data are divided into sub-groups, e.g., group 0 and group 1. Then, sub-models are trained using the respective sub-group data. Finally, after training the predictive system for each sub-group, edge descriptions of strong/weak relationships with path loadings are added to the figure.

Many feature-based TL algorithms have the function of data-dimension simplification. RF-TL is no exception. As mentioned in Section 3.2, in the data management step, items that do not have a strong ability to measure the model will be removed. Compared with other data-dimension reduction methods, the distinct advantage of using SEM is that the extracted dimensions are all meaningful in practice; e.g., Item 1 represents temperature and Item 2 represents weather. The meaningfulness of the item is a key point for the causal analysis. We use expert causal knowledge in the intervention step. "Knowledge" means something explainable; thus, it is impossible to do a further causal intervention in the succeeding steps without revealing the explanations of the data features.

3.4. Interventions on the source model

Intervention stimulates a model by artificial means, and the stimulation constrains the intervention item to be a constant state. Under the three assumptions of FPCI, when an analysis object is a group, causality can be represented by the expectation of the difference between the intervened state and original state, i.e., function (1). In addition, an intervention item, such as the intervening factor in the model in Fig. 5, is a measuring item that can be regarded as a characteristic for describing one of the factors of the model, e.g., factor 2 in Fig. 5. Here, we will give the following definitions:

Definition 3. *Intervention T*: *Classify objects into different sub-groups in accordance with the characteristic, i.e., the intervening factor. The intervening factor of each group is labeled by a constant number, such as* 1 *for the first group,* 2 *for the second group, etc.*

Definition 4. *The state* $E[Y_i(u)]$ *of the* i^{th} *group*: *The new prediction model trained by the data from sub-group i.*

Definition 5. *The original state* $E[Y_c(u)]$: *Assuming there are n sub-groups,* $E[Y_c(u)]$ *is the* $(n-1)^{th}$ *sub-group.*

There are a few caveats regarding these definitions. The first is about the division of the sub-groups. Data in D_S should be divided into sub-groups in accordance with the scale of the intervening factor in D_S. The division must have scale invariance. As in the baseball game example, if the temperature range of D_T is 5 °C and D_S is 15 °C, there will be three sub-groups, each having a scale of 5 °C. Second, RF-TL is based on the linear dependence between the reason and the result. Thus, the division of sub-groups is not random but in accordance with the increase or decrease in the intervention item. Furthermore, if the mean value of the intervention item of D_T is on the lower side of D_S, the intervention item of the sub-group is labeled in a descending way, i.e., winter is colder than summer, while if the mean value of the intervention item of D_T is on the higher side of D_S, they will be labeled in an ascending way. Here, if we suppose that D_T ranges from 0 °C to 5 °C and D_S ranges from 25 °C to the 40 °C, the 40 °C–35 °C sub-group can be labeled 1, the 35 °C–30 °C sub-group can be labeled 2, and the 30 °C–25 °C sub-group can be labeled 3. Third, the original state is needed for the causal analysis. The original state should be a group without any interventions. Nevertheless, it is difficult to find an ideal state without any intervention; thus, in practice, one of the intervened states is often chosen as the original one. We define a group with label $n-1$ as the original state for the convenience of evaluating its intervening scale relative to sub-group n, which is "nearest" D_T. The upper portion of Fig. 6 illustrates the intervention procedure.

After the intervention, the data in D_S are divided into sub-groups. Because the intervention item has been labeled with a constant number, which means the objects in the sub-group with such a label have the same attributes as the intervention item, the intervention item will no longer be a measuring item of the sub-models. After the intervention, SEM-like KGs are extracted using the data of each sub-group. The training procedure begins by preparing the data. The data features that are used as input for creating the i^{th} sub-prediction system P_i are those used by the prediction system before the intervention P. Unlike in the original D_S model, the items belonging to the same factor with the intervention item may be classified into another common factor in the EFA procedure because the intervention item is not used in the sub-models. It is also possible for the number of items or factors to decrease if the item does not have enough power to evaluate the system.

Table 1
Pseudo-code of RF-TL algorithm.

RF-TL: Transfer

Function EXECUTE TRANSFER (Q_{s1}, Q_{s2}, M_T, m, k, w)
  $M_T = \varnothing$
  for i in the range $(1, m \times (m-1) - k)$ do:
    $N_{s1}(edge_i, 0) \vee W_{s1}(edge_i, fl_{i\_s1}) \wedge S_{s2}(edge_i, fl_{i\_s2}) \Rightarrow M_T(edge_i)$
    $W_{s1}(edge_i, fl_{i\_s1}) \vee S_{s1}(edge_i, fl_{i\_s1}) \wedge N_{s2}(edge_i, 0) \Rightarrow M_T(\neg edge_i)$
    $S_{s1}(edge_i, fl_{i\_s1}) \wedge W_{s2}(edge_i, fl_{i\_s2}) \Rightarrow M_T(\neg edge_i)$
    $N_{s1}(edge_i, 0) \wedge W_{s2}(edge_i, fl_{i\_s2}) \Rightarrow$ LOADINGTRANSFER$(0, fl_{i_s2}, M_T, w)$
    $W_{s1}(edge_i, fl_{i\_s1}) \wedge W_{s2}(edge_i, fl_{i\_s2}) \Rightarrow$ LOADINGTRANSFER$(fl_{i_s1}, fl_{i_s2}, M_T, w)$
    $S_{s1}(edge_i, fl_{i\_s1}) \wedge S_{s2}(edge_i, fl_{i\_s2}) \Rightarrow$ LOADINGTRANSFER$(fl_{i_s1}, fl_{i_s2}, M_T, w)$
  end
  return M_T

RF-TL: Path-loading calculation

function LOADING TRANSFER ($fl_{i\_s1}, fl_{i\_s2}, M_T, w$)
  $fl_T = |fl_{i\_s2}| + (|fl_{i\_s2}| - |fl_{i\_s1}|) \times |w|$
  $|fl_T| \geq 0.3 \Rightarrow M_T(edge_i)$
  $|fl_T|. < 0.3 \Rightarrow M_T(\neg edge_i)$
  return M_T

Removal of items will affect the transfer process. The specific operations for handling this situation are discussed in the section about the transfer rules. However, the abstract concepts of the common factors should not be changed. Also, the meaning and number of latent factors should be the same in each sub-model; this is necessary for the following transfer procedure. If necessary, the common factors can be forced to be a certain number in accordance with the reference points. The procedure of creating sub-models is shown in the lower portion of Fig. 6. After creating the sub-predictive systems, the edge labels, i.e., weak/strong relationships with path loadings, are translated and added to the KGs.

3.5. Transferring knowledge graphs to the target domain

The purpose of RF-TL is to find suitable features for predicting the target in D_T through the transfer of the relationships of the D_S model. As mentioned in Section 2, a path loading (absolute value) ≥ 0.3 is the reference point for the predictive power of a factor. As a result, the transfer rules are defined for predicting the predictive power of the factors in the model of D_T. First-order logic programming (FOLP) [34] is used to create RF-TL, and the following pseudo-code shows the procedure. For a clear illustration, we have numbered the edges in the sub-models. As mentioned above, the number of latent factors remains the same in each sub-model. Assuming there are m factors in the model, if all the factors are connected to each other and the direction of the arrow is taken into account, there will be $m \times (m-1)$ edges. Also, as mentioned, if the path loading between nodes is extremely small, then no edge will be added to the KGs, i.e., $N_i(x,y,0)$. If $N_i(x,y,0)$ is true in all sub-models, this edge is considered to be useless for constructing the model. Thus, it is not necessary to input it to the transferring algorithm. In practice, most of the unnecessary data features are removed in the EFA step, and the remaining ones are classified into a few latent factors. As a result, the time cost of RF-TL is usually acceptable. Assuming there are k such edges, they will be ignored when labeling the edges. As a result, the labels from 1 to $m \times (m-1) - k$ are given to the (potential) edges of each model. The order does not matter, but it should be the same in each sub-model. The pseudo-code of RF-TL algorithm is shown in Table 1.

The inputs of the algorithm are Q_{s1}, Q_{s2}, M_T, m and w. Q_{s1} is the set of edges of the sub-model labeled $n-1$, and Q_{s2} is the set of edges of the sub-model labeled n. The edges are expressed using Functions (2)–(4). M_T is the transferred edges in D_T. m is the number of factors, and w is the transfer weight. The principal part of the transferring algorithm is performed according to FOLP. Specifically, whether an edge should be added to the KG of D_T is decided by comparing the "strengths" of the edges in the neighboring sub-models, model $n-1$ and model n. If the edge in model $n-1$ is weak or none and in model n is strong, then the edge is added to the target KG. In contrast, if the edge in model $n-1$ is strong or weak and in model n is none, then the edge is not added to the target KG. Similarly, if the edge in model $n-1$ is strong and in model n is weak, then the edge is not added to the target KG. Moreover, the other situations need the path loadings to be calculated using the path loading calculation algorithm.

In Section 3.2, the "reason" and "result" in the causal relationship used with RF-TL are defined as an intervening factor, such as "temperature" and a statistically expressible model's target, such as "attendance rate". The reason and result are assumed to have a linear relationship, so the causal relation can be expressed as

$$Result = \beta_r \times Reason \tag{5}$$

where β_r is the standard regression coefficient.
Furthermore, the transfer weight w is defined as

$$w = multiple \times \beta_r \tag{6}$$

Here, *multiple* depends on the "distance" between the highest (lowest) value of the intervention item $x_I^{(S)}$ in D_S and the lowest (highest) value of the intervention item $x_I^{(T)}$ in D_T. Although the sub-groups are

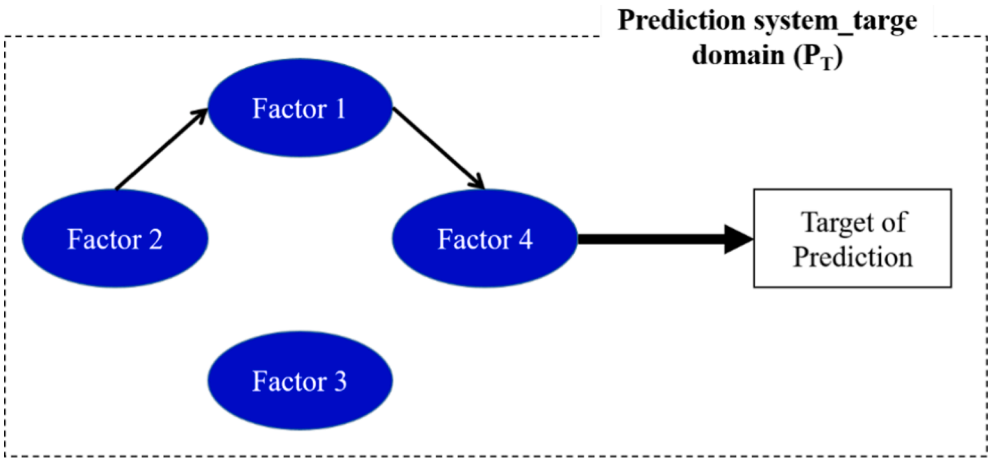


Fig. 7. The predictive system in D_T. After the transference, in the KG of D_T, only edges between Factors 1 and 2 and between Factor 1 and 4 are added. Factors 2, 3, and 4 are separated from each other. As Factor 4 directly points to the target of prediction and Factor 3 does not directly or indirectly connect to Factor 4, the data features that belong to Factor 3 will not be considered when the predicting system in D_T is constructed.

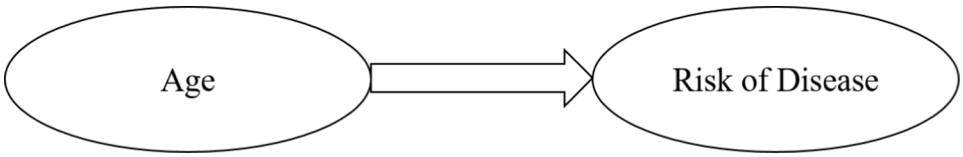


Fig. 8. Causal model of age increasing the risk of a particular disease, i.e., OSA and COVID-19.

Table 2
Hardware and software configurations of the experimental environment.

Hardware	
CPU	Core (TM) i7-7700HQ CPU @ 2.80 GHz 2.80 GHz
Memory	16 GB
Operation system, software and programing language	
OS	Windows 10 x64
Software	IBM SPSS Statistics v26
Programming Language	R-3.6.2
Key package in R	Lavaan-0.6–9, semPlot-1.1.2, GA-3.2.1,

Table 3
Parameter settings of RF-TL in the two experiments.

	OSA	**COVID-19**
m	5	2
k	16	1
w	0.975	0.994

divided up following the rules of the same scale of the intervention item in D_T, there may be a difference between the highest (lowest) value of it in D_S and the lowest (highest) value of it in D_T. As in the baseball game example, data in D_S range from 25 °C to 40 °C, but the data in D_T range from 0 °C to 5 °C There is a gap of 20 °C between the two domains. Here, *multiple* is used for filling the gap as shown in function (7). As mentioned in Section 3.4, we labeled the sub-groups according to the mean value of the intervention item of D_T and D_S. Similarly, when the mean value of the intervention item of D_T is higher than D_S, the distance is calculated by the lowest value of $x_I^{(T)}$ and the highest value of $x_I^{(S)}$, vice versa. The *scale* mentioned here corresponds to the range of $x_I^{(T)}$, which is also the basis for dividing sub-groups.

$$multiple = \begin{cases} 1 + \dfrac{\left|\min x_I^{(T)} - \max x_I^{(S)}\right|}{scale}, & if\ mean\left(x_I^{(T)}\right) > mean\left(x_I^{(S)}\right) \\ 1 + \dfrac{\left|\max x_I^{(T)} - \min x_I^{(S)}\right|}{scale}, & if\ mean\left(x_I^{(T)}\right) < mean\left(x_I^{(S)}\right) \end{cases} \quad (7)$$

The transfer weight w calculates the changing scale between the source domain and the target domain but does not consider the increase or decrease dependence that is decided by the label order of the sub-models mentioned in Section 3.4. Thus, in the transfer algorithm, the absolute value of w was used.

After calculating the path loadings, it is determined whether to add an edge to the target KG by comparing it with the threshold of 0.3.

3.6. Identifying data features for training models in the target domain

RF-TL returns a set of edges M_T, and all the edges are marked "strong". If there are nodes that do not connect to any other nodes, the items belonging to the nodes are unnecessary for the D_T model and will be removed. For example, for problem A, the final model in D_T is shown in Fig. 7.

As shown in Fig. 7, Factor 3 does not connect to any other factors in the prediction model after the transference. Thus, the items belonging to Factor 3 are removed and the items belonging to Factors 1, 2, and 4 are extracted for the prediction system in D_T.

Finally, the unlabeled target of prediction in D_T can be labeled by using a missing-data estimation method, such as the expectation–maximum (EM) algorithm.

4. Experiments

To evaluate the effectiveness of RF-TL, we conducted two experiments related to healthcare problems. One was on predicting obstructive sleep apnea (OSA). The other was on the prediction of ICU utilization in the COVID-19 pandemic. There is a common characteristic between these experimental cases, which is the higher the age of the patient is, the higher the risk will be [35,36,37]. Thus, expert causal knowledge in each case is the effect of age on the risk of disease. The causal model is shown in Fig. 8.

The hardware and software configurations of the experimental environment are shown in Table 2.

The primary development environment for the experiments is based on R. The essential package for SEM analysis is Lavaan, and the GA learns the structure of the graph. Finally, KGs are drawn using the semPlot package. Before building the structural model for SEM, we run the EFA in SPSS statistics software and the EM for predicting missing labels. Although the EFA and EM procedures can also be done in R, we took advantage of the user-friendly interaction of SPSS. The realization of the proposed algorithm is not limited to the configurations shown in Table 2. To the best of our knowledge, the mentioned packages can be used in other development environments, i.e., Python. The data sizes of the experiments were relatively small. When RF-TL is applied to big data, GPU-based packages can be used to speed up the calculation.

There are three parameters that need to be pre-set before running the RF-TL algorithm, m: number of nodes; k: number of "no branch in sub-models;" and w: transfer weight. The procedure for obtaining these parameters is shown in Section 3. In the respective experiments, these parameters were set as shown in Table 3.

4.1. Questionnaire diagnosis of obstructive sleep Apnea

OSA is a common sleep disorder. The most effective method of diagnosing OSA is using polysomnography with a peripheral capillary oxygen saturation test. However, it is expensive and difficult for people to use at home. Here, questionnaires are better than methods that require professional supervision as a means of diagnosing OSA in primary care and are self-diagnostic. There are many types of questionnaires containing numerous questions, such as the Quality of Life questionnaire, Epworth sleepiness scale, and Stop-Bang questionnaire [38,39]. We collected 60 items for predicting the risk of getting OSA from the self-rated questionnaires of the Sleep Heart Health Research dataset, which includes anthropometrics (6 items), health interviews (11 items), sleep habits, and quality (35 items), and 36-Item Short Form Survey (SF_36) questionnaires (8 calculated items). Additionally, an Apnea-Hypopnea Index (AHI) ≥ 5 is treated as undiagnosed OSA.

In the experimental dataset, there were a total of 3821 patients aged from 40 to 80. The patients in their 50 s and 60 s had labeled AHI data, and those in their 40 s and 70 s did not have any label. The tasks began with constructing an OSA-prediction model for patients in their 50 s and 60 s. Features for the young group (40 s) and old group (70 s) were transferred from the 50 s ~ 60 s model. D_{SO} was the source domain that included the features X_{SO} (in their 50 s and 60 s with the label of AHI),

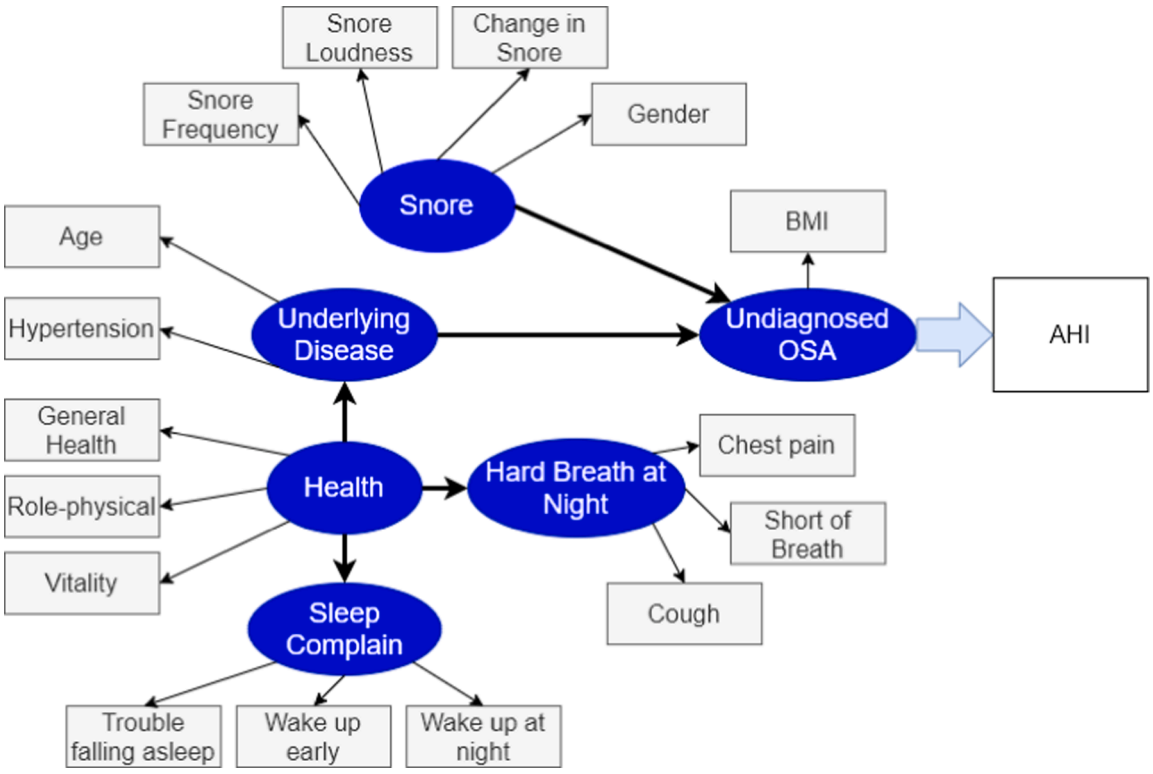


Fig. 9. Causal model for predicting OSA in the source domain.

D_{TO1} was the target domain that included the features X_{TO1} (in their 40 s without the label of AHI), and D_{TO2} was the target domain that included the features X_{TO2} (in their 70 s without the label of AHI). There were two tasks. T_{TO1} was to extract the data feature for predicting OSA in $D_{TO1,}$ and T_{TO2} was to extract the data feature for predicting OSA in D_{TO2}.

Fig. 9 shows the constructed model for predicting AHI in D_{SO}, which consists of 16 questionnaire item variables that are classified into six factors. Age was one of the measuring items for the factor "underlying disease." The age intervention produced two groups. One was a sub-group of patients in their 50 s labeled Younger, and the other was a sub-group of patients in their 60 s labeled Older. Two sub-models were trained using the 16 items with the corresponding data in each sub-group. The trained sub-models are shown in Fig. 10.

As shown in Fig. 10, the age intervention removed the factor "underlying disease" from the model, and classified the Hypertension item into the factor "undiagnosed OSA." This re-classification is reasonable and will not influence the result of the transfer. The factor loadings are marked on the path and have been translated into "weak" or "strong" labels.

Next, RF-TL was used to transfer the knowledge network from D_{SO} to D_{TO}. As mentioned above, there were two transfer tasks. One was to transfer from D_{SO} to $D_{TO1,}$ (Young). Here, the Younger sub-model was labeled 1 and Older sub-model was labeled 2. The other task was to transfer from D_{SO} to $D_{TO2,}$ (Old). Here, the Younger sub-model was labeled 2 and Older sub-model was labeled 1. Next, the counterfactual inference was performed on the basis of causal knowledge. For example, as shown in Fig. 10(a), the relationship between factor "health" and factor "snore" is strong with a factor loading of −0.46. In Fig. 10(b), the relationship is still strong but with a factor loading of −0.31. Additionally, the "reason" we used in this example is "age". Thus, from Fig. 10, we can obtain the following information:

> "In the prediction system of AHI, as age increases (decreases), the relationship between Health and Snore becomes weaker (stronger)."

Furthermore, the counterfactual inference yielded,

> "If age is older (younger), then the relationship between health and snore is weaker (stronger)."

The counterfactual inference of the RF-TL algorithm is quantified depending on which the relationship between factors in the target domain is predicted (such as the weak relationship between "health" and "snore" in T_{TO2} as shown in Fig. 10(b)). The transferred models for T_{TO1} and T_{TO2} are shown in Fig. 11.

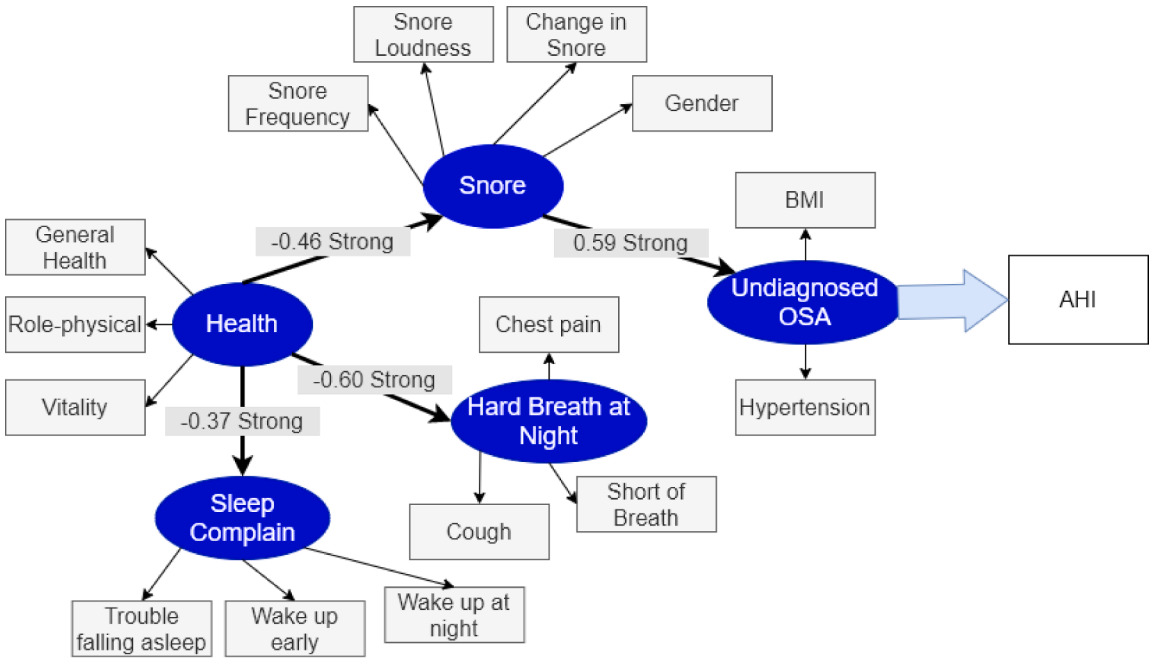
For the transferred model of T_{TO1} (Young), all the relations in the model were marked as strong and connected. As well as the items shown in the model, we used age as the expert knowledge. As a result, age was found to be the ground truth that changes OSA and that should be used as one of the data features for predicting AHI. Thus, there were 15 items, plus the item age, which was used for predicting AHI for the objects aged in their 40 s.

Different from the model for T_{TO1}, the relationship starting from Health and pointing to Snoring and Sleep Complaints was "Weak" for D_{TO2}. Thus, an edge was not added to the graph, and the model was divided into two parts: one part with the factors "Health", "Sleep Complaints", and "Difficult to Breathe at Night" and one part for predicting AHI. Only the six items contained in the factors "Snoring and Undiagnosed OSA" and the age item were used for predicting AHI of T_{TO2}.

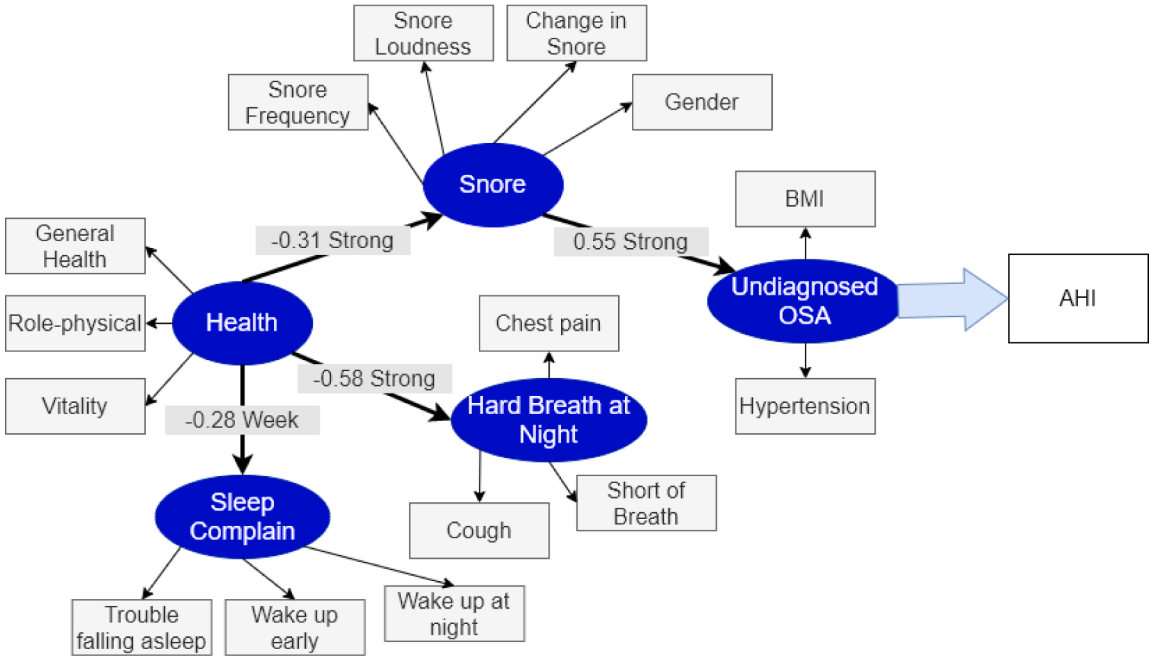
To evaluate the effectiveness of the feature transfer, the EM algorithm was used for predicting the label for the target domains. As mentioned above, we extracted 16 items for T_{TO1} and 7 items for T_{TO2} from the 60 items originally collected. We used the EM algorithm for labelling AHI in D_{TO1} and D_{TO2} with 60 items, 16 items, and 7 items. We used accuracy and F1_score as evaluation indexes. Table 4 lists the results.

The prediction results indicate that the use of 16 items in D_{TO1} resulted in the highest accuracy and F1_score and that 7 was the most suitable number of extracted items for predicting the AHI of the old group.

The CORAL algorithm and TCA algorithm are two commonly used TL algorithms for transferring features from the source to the target domain



(a) Younger sub-model (in their 50s)

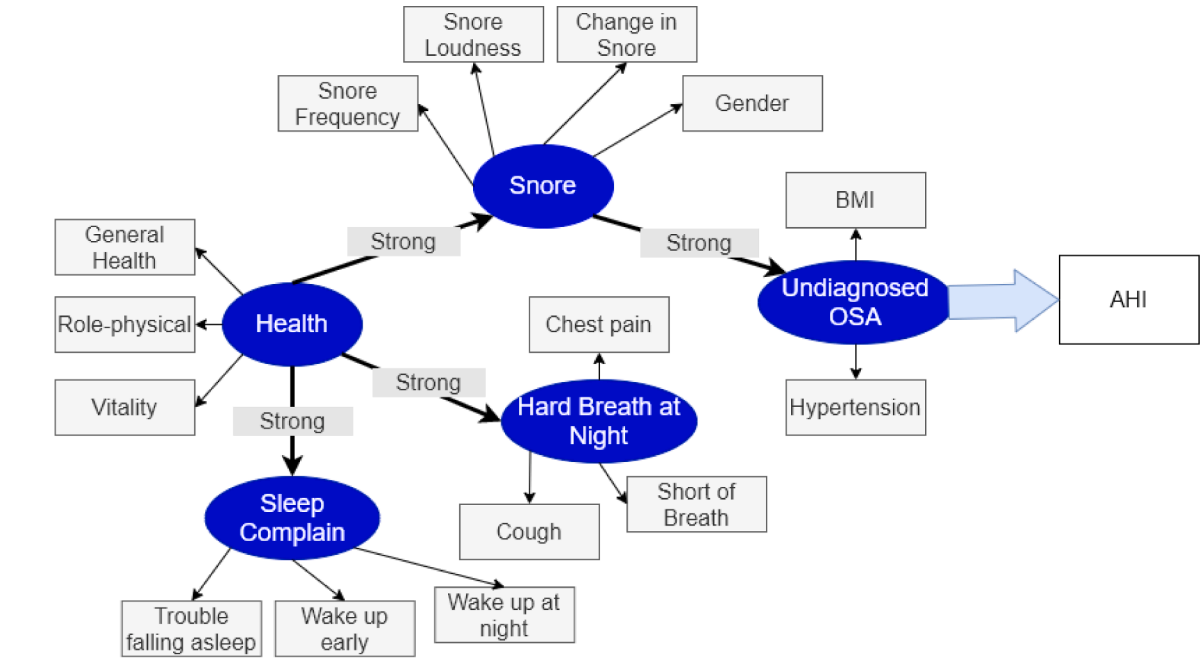


(b) Older sub-model (in their 60s)

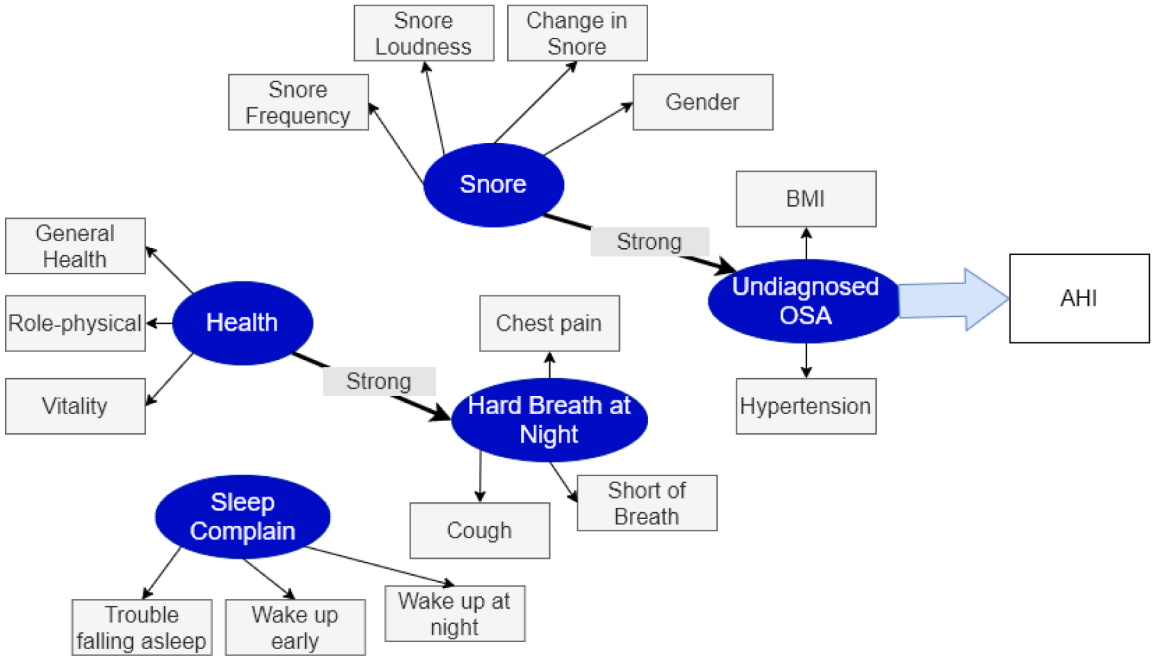
Fig. 10. Sub-models for predicting OSA in the divided source domains.

with no labels. We compared these two algorithms with RF-TL. For TCA, it is necessary to determine the number of previously transferred features; thus, 16 is given to T_{TO1} and 7 is given to the T_{TO2} to maintain consistency with RT-TL. On the other hand, CORAL does not need to define the number of previously transferred features and transferred 59 items from the source domain to both target domains. The number of features it extracted was much greater than the number extracted by RF-TL, which highlights the advantage of RF-TL. The EM algorithm was used again for predicting AHI by using the items extracted with TCA and CORAL, and the results were compared with those of RF-TL. Table 5 lists the results.

For T_{TO1}, RF-TL had the highest accuracy and F1_score. For T_{TO2}, although the accuracy of TCA was higher, the precision of negative (AHI is labeled as 0) was zero, so there was no F1_score and it failed to make a



(a) Transferred model for T_{TO1}



(b) Transferred model for T_{TO2}

Fig. 11. Transferred models for predicting OSA in the two different target domains.

prediction. The accuracy of RF-TL was higher than that of CORAL. The F1_score was a little lower due to the unbalanced number of objects contained in the negative and positive groups. Also, there were 59 items used for CORAL and only 7 items used for RF-TL. These results show that RF-TL outperformed CORAL.

4.2. *ICU-candidate prediction for COVID-19 patients*

The novel coronavirus started spreading across the world in early 2020. Millions of people have been infected, and the number is still increasing. Because of the large number of patients, medical collapse threatens many countries. Predicting severe cases requiring an intensive care unit (ICU) is an important task. The "COVID-19 - Clinical Data to assess diagnosis" dataset has been published online [40]; it contains 189

Table 4
Comparison of OSA prediction using different numbers of items.

D_{TO}	Young (D_{TO1})		Old (D_{TO2})	
No.	**Accuracy**	**F1_score**	**Accuracy**	**F1_score**
60 items	72.88%	0.7319	75.16%	0.5669
16 items	**74.53%**	**0.7519**	76.14%	0.5801
7 items	73.71%	0.7441	**76.41%**	**0.5804**

Table 5
Comparison of results with TCA and CORAL for predicting OSA.

D_{TO}	Young (D_{TO1})		Old (D_{TO2})	
Methods	**Accuracy**	**F1_score**	**Accuracy**	**F1_score**
TCA	55.28%	0.6098	77.49%	–
CORAL	56.84%	0.6573	72.74%	0.5934
RF-TL	**74.53%**	**0.7519**	**76.41%**	**0.5804**

Table 6
Information on patients infected by COVID-19.

Age group	No. of patients	ICU-positive rate [%]
20 s and 30 s	113	36.28
40 s and 50 s	110	43.64
60 s and 70 s	108	55.56
80 s and 90 s	218	62.63

items, including the ICU item (0 for No, 1 for Yes) collected from the patients diagnosed with COVID-19. There are 430 objects with no missing items of the patients ranging in age from their 20 s to 90 s. The distribution of the age groups and the ICU-positive rate are listed in Table 6.

Note that the data were collected before the mutated virus started to spread. The example only considered age as the factor in the counterfactual inference. Note as well that the current situation of viral spread is different due to mutations (such as widespread transmission of the mutated virus among young people), which may cause differences from the results of this example. RF-TL only considered single-factor causality. The limitations of this point will be explained in the discussion section.

The ICU-positive rate has a positive correlation with age. Also, in accordance with current knowledge, age is one of the factors of infection and severe cases [41], which conforms to the age-disease causal model shown in Fig. 8. We assumed that only the data of the 40 s–70 s age groups were labeled with ICU tags (DSI) and that the two target domains D_{TI1} of the 20 s and 30 s groups and D_{TI2} of the 80 s and 90 s groups did not have ICU tags. RF-TL was used for transferring data features from D_{SI} to D_{TI1} and D_{TI2}. The two tasks, T_{TI1} and T_{TI2}, aimed at extracting suitable data items for predicting ICU candidates in D_{TI1} and D_{TI2}.

A prediction model was first constructed for D_{SI}, as shown in Fig. 12.

The notation V_O2 denotes the partial pressure of venous oxygen (minimum); V_CO2 denotes the partial pressure of venous carbon dioxide (maximum); V_SATO2 denotes blood oxygen saturation (mean); BP denotes diastolic blood pressure (range); RRM denotes respiratory rate (mean); and RRD denotes respiratory rate (range/median).

Although there were 189 items available, only 8 items were extracted for predicting whether a patient needs to be sent to the ICU. An intervention was conducted on the age factor. Different from the OSA case, age in the ICU model is a factor, not an item. Thus, the invention procedure removes the age factor from the model and divides the source domain into two sub-domains; one containing patients in their 40 s and 50 s and the other containing patients in their 60 s and 70 s. The sub-models are shown in Fig. 13.

The rules were used to transfer the models to the target domains. The KGs for T_{TI1} and T_{TI2} are shown in Fig. 14.

From the information in Fig. 14, only the 3 items belonging to the Potential ICU factor and the age item were used for T_{TI1}. Six items with age, a total of 7 data features were used for T_{TI2}. The EM algorithm was used for labeling the ICU data. Table 7 compares the prediction results for T_{TI1} and T_{TI2} with 189, 7, and 4 items.

As expected, the 4 items of T_{TI1} and 7 items of T_{TI2} yielded the highest prediction performance.

Similarly, we compared the predictions with those of TCA and CORAL. For the COVID-19 ICU case, 186 items were transferred from the source domain to the two target domains with CORAL and items were fixed for the younger domain and 7 items for the older domain with TCA. The results are listed in Table 8. RF-TL performed the best in each target domain.

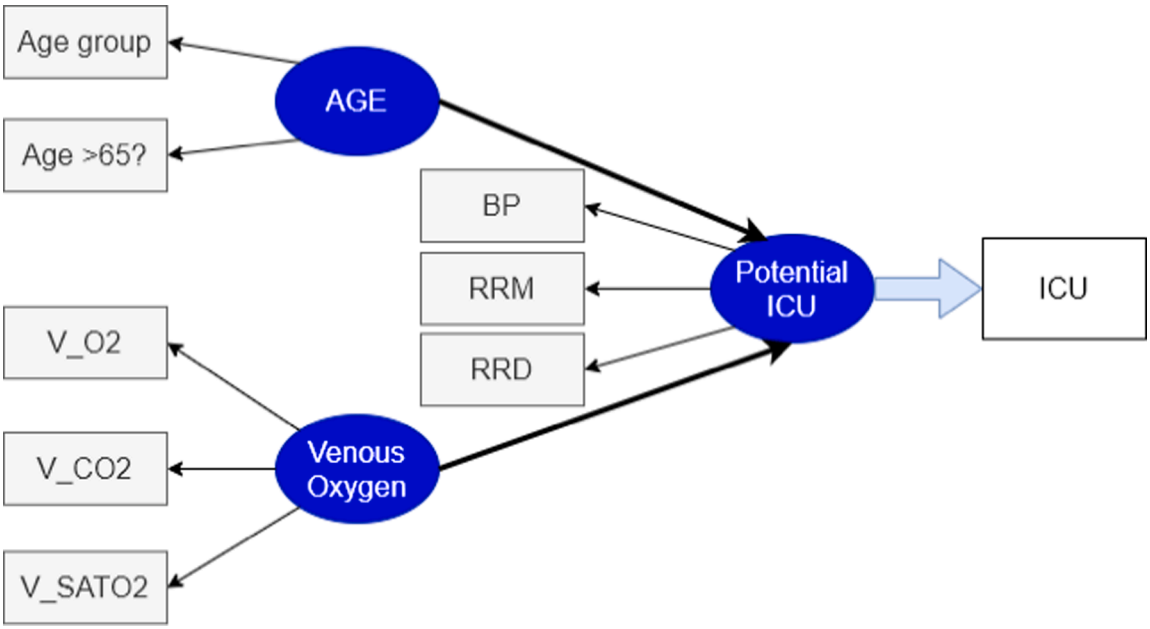
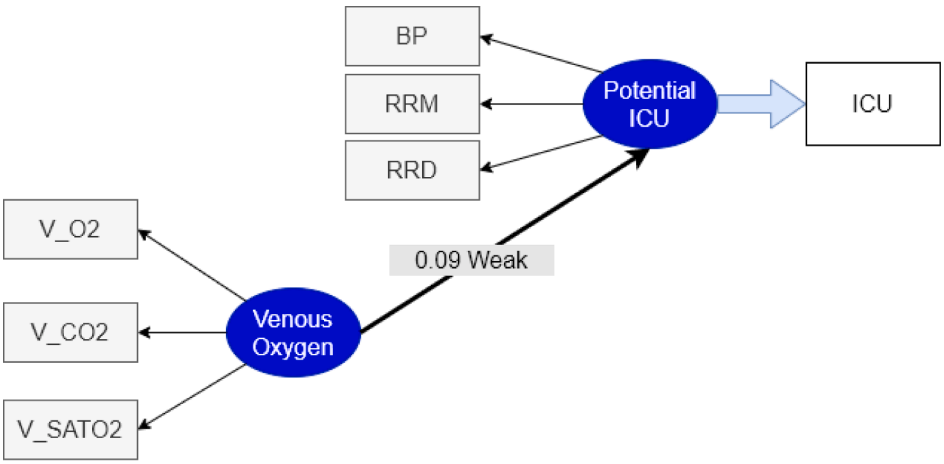
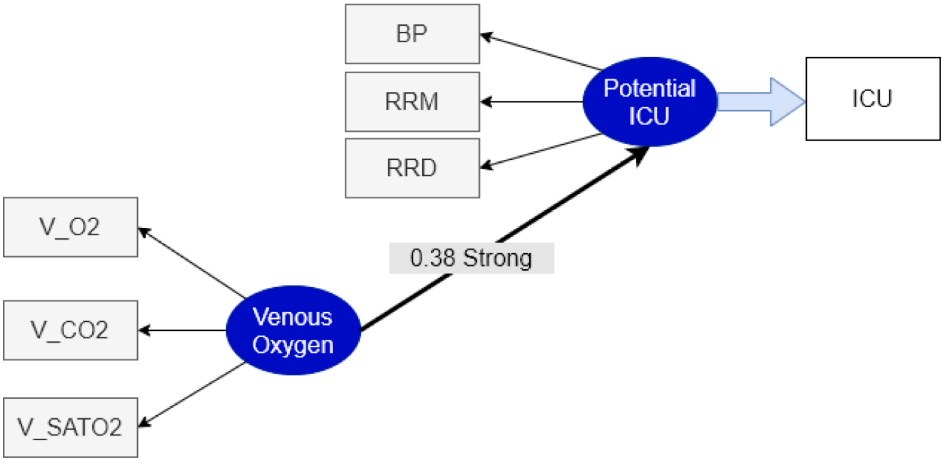


Fig. 12. Causal model for ICU-candidate prediction in the source domain.



(a) Sub-model for younger group (in their 40s and 50s)



(b) Sub-model for older group (in their 60s and 70s)

Fig. 13. Sub-models for ICU-candidate prediction in the divided source domains.

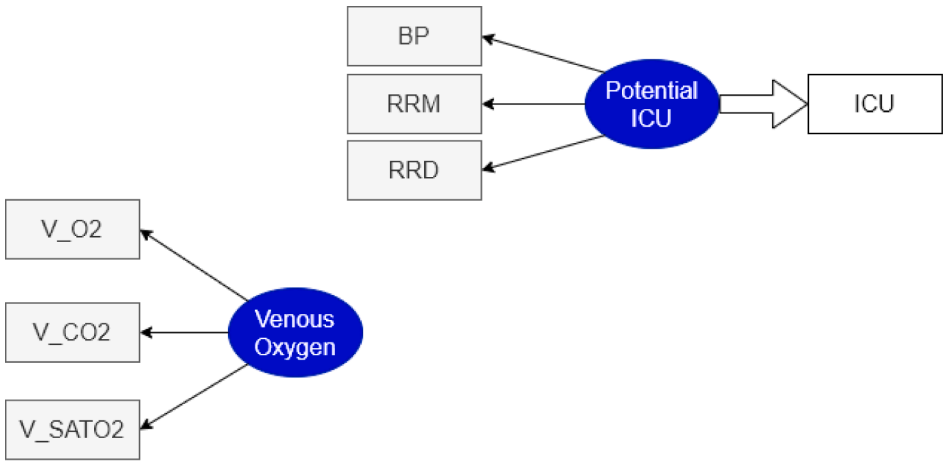
5. Discussion

The transferred element of RF-TL is the relationship in the model, which is different from other TL algorithms. Apart from accuracy, researchers of ML technologies are beginning to focus their attention on the inferring logic inside the model. Their goal is to build ML models with the ability to interpret human cognitive and reasoning processes. KGs are excellent tools for showing human knowledge networks in which domain experts' inference logic can be demonstrated. Relational TL algorithms are applications of KGs. For relational TL, only by clarifying the learning structure of the source-domain model can the relationships be transferred to target domains.

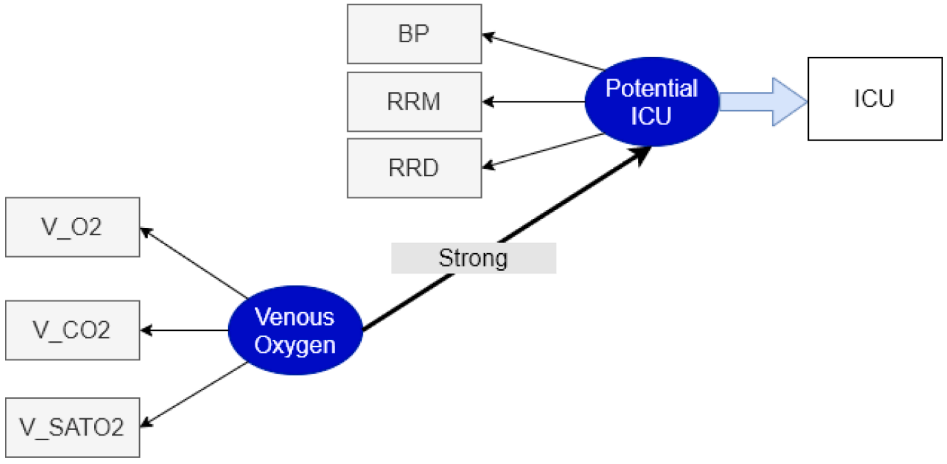
A causal relationship is a higher level of statistical dependency between two data items and is ground truth based on expert knowledge or experience. The reason and the result in a causal model are correlated with each other. However, the causality between them cannot be determined only by clarifying the correlation between two items. Causality needs a "time order". That is, the reason occurs before the result, and a change in the reason inevitably causes a corresponding change in the result. In contrast, correlation is only an expression of the data at a certain time point and it does not express the time order. Future prediction needs to clarify the development of one thing along with the time stream. This is why counterfactual inference can be made only in accordance with the causal relationship.

Transferring the information from a known domain to an unknown domain can be treated as a prediction; thus, the level of statistical dependency is not sufficient. Traditional TL algorithms, such as TCA and CORAL, only take into account the statistical relationships among data features. The two algorithms do not perform well because of the non-significant difference in the data-feature distribution between the source domain and the target domain. Nevertheless, RF-TL uses causality to direct the transfer procedure by predicting how the relationships between data feature changes in the source domain. The relations among the features are considered, and the inference is performed in accordance with explainable human causal knowledge. As a result, good performance can be obtained in practical applications.

However, the two experiments had limited data sizes, so the calculating time was not long. Two parts of RF-TL take up most of the calculation time. One is the comparison of the edges between sub-models. The more edges are added, the higher the time cost becomes. To deal with this problem, RF-TL uses a pre-pruning step before transferring. As shown in the OSA experiment, 16 edges are removed from 20 edges before the transference steps. The other time-consuming step is the GA procedure for identifying the structure of KGs. Referring to SEM-



(a) KG of D_{TI1}



(b) KG of D_{TI2}

Fig. 14. Transferred models for ICU-candidate prediction in the two different target domains.

Table 7
Comparison of ICU prediction using different numbers of items.

D_{TO}	Young (D_{TI1})		Old (D_{TI2})	
No.	**Accuracy [%]**	**F1_score**	**Accuracy [%]**	**F1_score**
189 items	84.96	0.8353	79.80	0.8018
7 items	84.07	0.7519	**88.98**	**0.8898**
4 items	**89.38**	**0.8895**	77.78	0.7687

Table 8
Comparison of results of predicting OSA with TCA and CORAL.

D_{TI}	Young (D_{TI1})		Old (D_{TI2})	
Algorithms	**Accuracy**	**F1_score**	**Accuracy**	**F1_score**
TCA	36.28%	–	62.63%	–
CORAL	81.42%	0.8124	80.81%	0.8198
RF-TL	**89.38%**	**0.8895**	**88.89%**	**0.8898**

EML, the proposed SEM-like KGs conducts a two-step GA. In step 1, correlations are estimated between each pair of nodes. It can be easily performed in SEM by adding double-direction arrows to all nodes. Edges connecting nodes with a correlation coefficient higher than 0.1 are labeled. Then, the labeled edges that are suggested solutions to be input to the GA are constrained to be "1". The Goodness of Fit (GoF) indexes are used as fitting functions in the GA, and the suggested edges let GA iterations start from a relatively high GoF which helps to reduce the number of iterations and save calculation time. Although various factors that influence time costs are considered, RF-TL needs a further test to determine its effectiveness and feasibility. On the other hand, the causal model used in RF-TL is a single-factor causality. In other words, only parts of the causal structure are taken into account. For a practical case, one result is usually linked by multiple reasons. As in the COVID-19 example, except for age, mutation of the virus would be another factor influencing the prediction of the severe-case rate. When considering multiple factors in a causal model, the degree of influence of each factor should be weighted accordingly.

Additionally, except for the medical cases mentioned in this study, the intelligent data-driven knowledge networks are useful in industrial practice, such as the design and application of the Smart Product Service System (Smart PSS) [41,42,43]. In order to meet users' diverse and variable expectations, the industrial applications need to identify important knowledge and relationships from a large amount of collected information and cope with the change in the collected data [44,45]. The

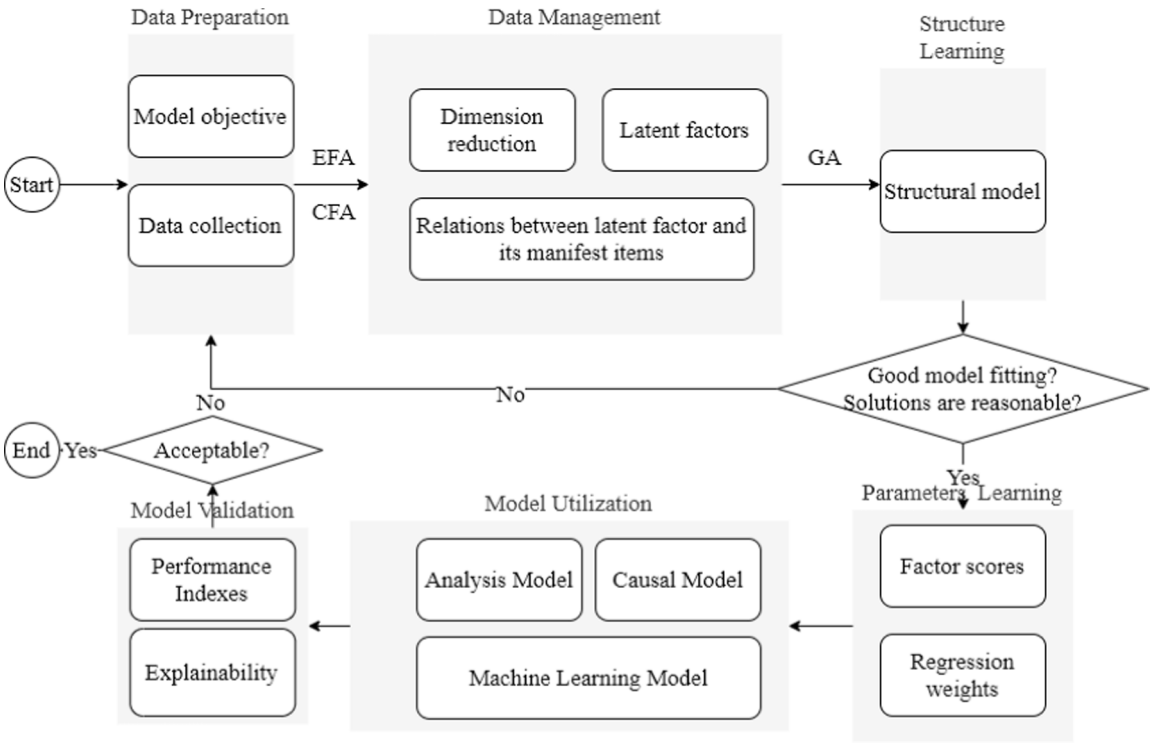


Fig. A1. Overall structure of SEM-EML.

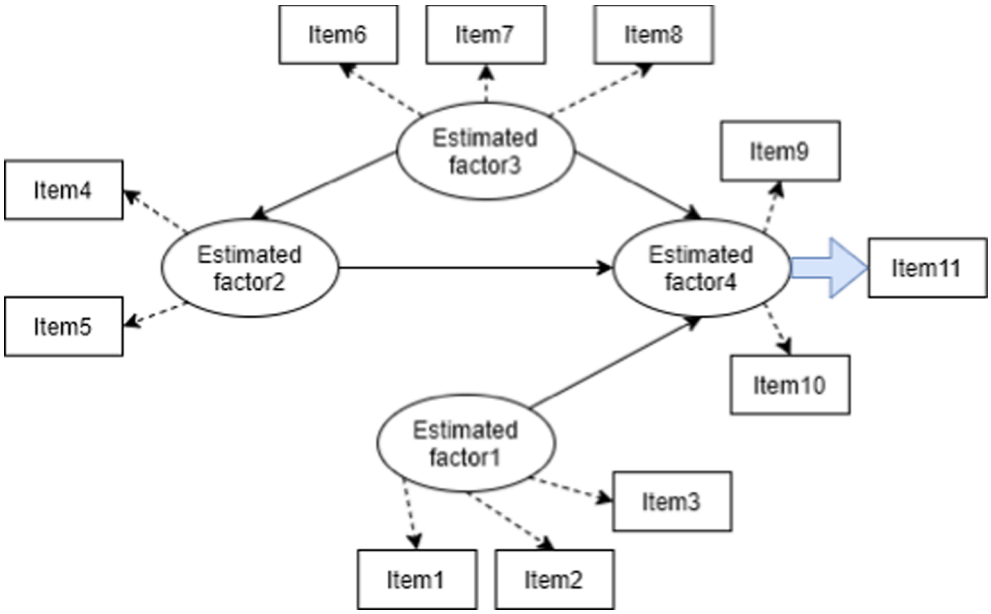


Fig. A2. Prediction model obtained by SEM-EML.

proposed RF-TL is helpful to accurately identify the key knowledge from data as well as the causal relationships among the concepts of the extracted information. Also, as a transfer learning technique, the utilization of RF-TL can be extended to other engineering practices for predicting the knowledge networks in an unknown domain even without sufficient data, e.g., designing the service system according to the anticipated users' diversified expectations. We will do it in our future work.

6. Conclusions

In accordance with the directions of causal analysis and counterfactual inference, RF-TL transfers relationships between data features from a source domain to the target domain. Feature extraction is then conducted in accordance with the information in the transferred KG. Because RF-TL takes into account links between different data items and the prediction function of causality, it performs better in practical cases compared with other TL algorithms. RF-TL is based on linear relations and single-factor causality. In the future, we will extend it to the non-linear domain and develop a multiple-factors causal model.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

Explainable Machine Learning Model based on Structural Equation Modeling (SEM-EML)

SEM-EML is based on SEM. The procedure of the SEM-EML contains six steps: data preparation, data management, structure learning, parameter learning, model utilization, and model validation. The overall procedure is shown in Fig. A1.

For a machine learning problem B, m data features are collected. After the data management and structure learning steps, $n(n \leq m)$ features are identified as necessary for constructing a prediction model, in which n features are classified into p factors. For the sake of illustration, we will assume that, for problem B, 15 data features are collected, from which 11 features remain for constructing the prediction model, and item 11 is the target of prediction. The model gotten by SEM-EML is shown in Fig. A2.

SEM-EML can simplify the data features and classify the features it extracts into common factors. Linear regression is used for estimating the factor scores of the common factors. Equation (A1) is used to compute the estimated factor score of the i^{th} factor, i.e.,NFS_i,

$$NFS_i = N\beta_i + N\omega_{i_1} \times item_1 + \cdots + N\omega_{i_j} \times item_j + \cdots + N\omega_{i_{n-1}} \times item_{m-1} \tag{A1}$$

Here, we assume that m items, including the predicted target, are extracted and the m^{th} item is the target.$N\omega_{i_j}$ is the standard regression weight of $item_j$ for the i^{th} factor, and $N\beta_i$ is a constant number. The model separates the target item from the other ones by using a regression method. The prediction procedure is conducted only on NFS_t, where t is the index of the factor that directly connects to the target item, e.g., Estimated factor 4 in Fig. A2. Unsupervised clustering can be used to classify NFS_t and label the target. Although the other factors are not directly used to predict the target, they are abstract descriptions of the items used for making predictions, and they make up the knowledge network for the prediction.

References

[1] L. Torrey, J. Shavlik, Transfer learning. In Handbook of research on machine learning applications and trends: algorithms, methods, and techniques, IGI global (2010) 242–264.
[2] K. Weiss, T.M. Khoshgoftaar, D. Wang, A survey of transfer learning, J. Big data 3 (1) (2016) 9.
[3] S.J. Pan, I.W. Tsang, J.T. Kwok, Q. Yang, Domain adaptation via transfer component analysis, IEEE Trans. Neural Networks 22 (2) (2011) 199–210.
[4] B. Sun, J. Feng, K. Saenko, Return of frustratingly easy domain adaptation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 30(1), 2016, March.
[5] J. Wang, Y. Chen, S. Hao, W. Feng, Z. Shen, Balanced distribution adaptation for transfer learning, in: 2017 IEEE international conference on data mining (ICDM), 2017, pp. 1129–1134.
[6] J. Blitzer, R. McDonald, F. Pereira, July). Domain adaptation with structural correspondence learning, in: In Proceedings of the 2006 conference on empirical methods in natural language processing, 2006, pp. 120–128.
[7] J. Wang, Y. Chen, W. Feng, H. Yu, M. Huang, Q. Yang, Transfer learning with dynamic distribution adaptation, ACM Trans. Intell. Syst. Technol. (TIST) 11 (1) (2020) 1–25.
[8] IBM Cloud Education (2021.4). Knowledge Graph. Retrieved from https://www.ibm.com/cloud/learn/knowledge-graph.
[9] A.T. Bimba, N. Idris, A. Al-Hunaiyyan, R.B. Mahmud, A. Abdelaziz, S. Khan, V. Chang, Towards knowledge modeling and manipulation technologies: A survey, Int. J. Inf. Manage. 36 (6) (2016) 857–871.
[10] H. Chen, X. Luo, An automatic literature knowledge graph and reasoning network modeling framework based on ontology and natural language processing, Adv. Eng. Inf. 42 (2019) 100959, https://doi.org/10.1016/j.aei.2019.100959.
[11] L. Li, P. Wang, J. Yan, Y. Wang, S. Li, J. Jiang, Z. Sun, B. Tang, T.-H. Chang, S. Wang, Y. Liu, Real-world data medical knowledge graph: construction and applications, Artif. Intell. Med. 103 (2020) 101817, https://doi.org/10.1016/j.artmed.2020.101817.
[12] B. Cheng, Y. Zhang, D. Cai, W. Qiu, D. Shi, Construction of traditional Chinese medicine knowledge graph using data mining and expert knowledge, in: 2018 International Conference on Network Infrastructure and Digital Content (IC-NIDC), 2018, pp. 209–213.
[13] D. Shi, T. Wang, H. Xing, H. Xu, A learning path recommendation model based on a multidimensional knowledge graph framework for e-learning, Knowl.-Based Syst. 195 (2020) 105618, https://doi.org/10.1016/j.knosys.2020.105618.
[14] G. Peng, H. Wang, H. Zhang, K. Huang, A hypernetwork-based approach to collaborative retrieval and reasoning of engineering design knowledge, Adv. Eng. Inf. 42 (2019) 100956, https://doi.org/10.1016/j.aei.2019.100956.
[15] M. Rotmensch, Y. Halpern, A. Tlimat, S. Horng, D. Sontag, Learning a health knowledge graph from electronic medical records, Sci. Rep. 7 (1) (2017) 1–11.
[16] J. Li, T. Sawaragi, Y. Horiguchi, Introduce structural equation modelling to machine learning problems for building an explainable and persuasive model, SICE J. Control, Measurement Syst. Integrat. 14 (2) (2021) 67–79.
[17] P.G. Omran, K. Wang, Z. Wang, December). Transfer learning in probabilistic logic models, in: Australasian Joint Conference on Artificial Intelligence, Springer, Cham, 2016, pp. 378–389.
[18] R. Kumaraswamy, P. Odom, K. Kersting, D. Leake, S. Natarajan, Transfer learning via relational type matching, in: 2015 IEEE International Conference on Data Mining, 2015, pp. 811–816.
[19] R. Kumaraswamy, N. Ramanan, P. Odom, S. Natarajan, Interactive Transfer Learning in Relational Domains, KI-Künstliche Intelligenz 34 (2) (2020) 181–192.
[20] F. Wang, Z. Jiang, X. Li, G. Li, Cognitive factors of the transfer of empirical engineering knowledge: A behavioral and fNIRS study, Adv. Eng. Inf. 47 (2021) 101207, https://doi.org/10.1016/j.aei.2020.101207.
[21] D. Gentner, L. Smith, Analogical reasoning, in: Encyclopedia of Human Behavior, Elsevier, 2012, pp. 130–136, https://doi.org/10.1016/B978-0-12-375000-6.00022-7.
[22] M. Rojas-Carulla, B. Schölkopf, R. Turner, J. Peters, Invariant models for causal transfer learning, J. Mach. Learn. Res. 19 (1) (2018) 1309–1342.
[23] X. Chen, S. Jia, Y. Xiang, A review: Knowledge reasoning over knowledge graph, Expert Syst. Appl. 141 (2020) 112948, https://doi.org/10.1016/j.eswa.2019.112948.
[24] M. Kim, F. Lu, V.V. Raghavan, Automatic construction of rule-based trees for conceptual retrieval, in: Proceedings Seventh International Symposium on String Processing and Information Retrieval. SPIRE 2000. IEEE, 2000, September, pp. 153-161.
[25] M. Tenorth, M. Beetz, KnowRob: A knowledge processing infrastructure for cognition-enabled robots, Int. J. Robotics Res. 32 (5) (2013) 566–590.
[26] A. Gebharter (Ed.), Causal Nets, Interventionism, and Mechanisms, Springer International Publishing, Cham, 2017.
[27] G.W. Imbens, D.B. Rubin, Causal inference in statistics, social, and biomedical sciences, Cambridge University Press, 2015.
[28] G.W. Imbens, D.B. Rubin, Rubin causal model, in: S.N. Durlauf, L.E. Blume (Eds.), Microeconometrics, Palgrave Macmillan UK, London, 2010, pp. 229–241, https://doi.org/10.1057/9780230280816_28.
[29] L. Tao, W. Cichen, L. Huakang, Development and construction of knowledge graph, J. Nanjing Univ. Sci. Technol. (2017).
[30] A. Singhal, Introducing the knowledge graph: things, not strings, Official Google blog 5 (2012).
[31] J.Z. Pan, G. Vetere, J.M. Gomez-Perez, H. Wu, Exploiting linked data and knowledge graphs in large organizations, Springer, Heidelberg, 2017, p. 281.
[32] J.J. Hox, C.J.M. Maas, The accuracy of multilevel structural equation modeling with pseudo balanced groups and small samples, Struct. Equ. Model. 8 (2) (2001) 157–174.
[33] H. Steinmetz, R. Isidor, N. Baeuerle, April). Testing the circular structure of human values: A meta-analytical structural equation modeling approach, Survey Res. Methods 6 (1) (2012) 61–75.
[34] N. Lavrac, S. Dzeroski, Inductive Logic Programming, in: WLP, pp. 146-160, 1994.
[35] I.E. Gabbay, P. Lavie, Age-and gender-related characteristics of obstructive sleep apnea, Sleep Breathing 16 (2) (2012) 453–460.
[36] J. Krieger, E. Sforza, A.n. Boudewijns, M. Zamagni, C. Petiau, Respiratory effort during obstructive sleep apnea: the role of Age and sleep state, Chest 112 (4) (1997) 875–884.
[37] L. Ayalon, S. Ancoli-Israel, S.P.A. Drummond, Obstructive sleep apnea and Age: a double insult to brain function? Am. J. Respir. Crit. Care Med. 182 (3) (2010) 413–419.

[38] G.Q. Zhang, L. Cui, R. Mueller, et al., The National Sleep Research Resource: towards a sleep data commons, J. Am. Med. Inform. Assoc. 25 (10) (2018) 1351–1358.
[39] S.F. Quan, B.V. Howard, C. Iber, et al., The sleep heart health study: design, rationale, and methods, Sleep 20 (12) (1997) 1077–1085.
[40] Hospital Sírio-Libanês (2020.7). COVID-19 - Clinical data to assess diagnosis. Retrieved from https://www.kaggle.com/S%C3%ADrio-Libanes/covid19.
[41] J.T. Wu, K. Leung, M. Bushman, N. Kishore, R. Niehus, P.M. de Salazar, G. M. Leung, Estimating the clinical severity of COVID-19 from the transmission dynamics in Wuhan, China, Nat. Med. 26 (4) (2020) 506–510.
[42] X. Li, C.H. Chen, P. Zheng, Z. Wang, Z. Jiang, Z. Jiang, A knowledge graph-aided concept–knowledge approach for evolutionary smart product–service system development, J. Mech. Des. 142 (10) (2020).
[43] Z. Wang, C.-H. Chen, P. Zheng, X. Li, L.P. Khoo, A novel data-driven graph-based requirement elicitation framework in the smart product-service system context, Adv. Eng. Inf. 42 (2019) 100983, https://doi.org/10.1016/j.aei.2019.100983.
[44] Z. Wang, C.-H. Chen, P. Zheng, X. Li, L.P. Khoo, A graph-based context-aware requirement elicitation approach in smart product-service systems, Int. J. Prod. Res. 59 (2) (2021) 635–651.
[45] X. Li, C.-H. Chen, P. Zheng, Z. Jiang, L. Wang, A context-aware diversity-oriented knowledge recommendation approach for smart engineering solution design, Knowl.-Based Syst. 215 (2021) 106739, https://doi.org/10.1016/j.knosys.2021.106739.