Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Artificial intelligence-based decision support model for new drug development planning

Ye Lim Jung a,*, Hyoung Sun Yoo a,b,*, JeeNa Hwang a

a Division of Data Analysis, Korea Institute of Science and Technology Information (KISTI), Seoul 02456, Republic of Korea
b Science and Technology Management Policy, University of Science and Technology, Daejeon 34113, Republic of Korea

A R T I C L E   I N F O

Keywords:
Decision support model
Hybrid recommender system
Drug development recommendation
Pharmaceutical portfolio management
COVID-19 vaccine development prediction

A B S T R A C T

New drug development guarantees a very high return on success, but the success rate is extremely low. Pharmaceutical companies have attempted to use various strategies to increase the success rate of drug development, but this goal has been difficult to achieve. In this study, we developed a model that can guide effective decision-making at the planning stage of new drug development by leveraging machine learning. The Drug Development Recommendation (DDR) model, we present here, is a hybrid model for recommending and/or predicting drug groups suitable for development by individual pharmaceutical companies. It combines association rule learning, collaborative filtering, and content-based filtering approaches for enterprise-customized recommendations. In the case of content-based filtering applying a random forest classification algorithm, the accuracy and area under curve were 78% and 0.74, respectively. In particular, the DDR model was applied to predict the success probability of companies developing Coronavirus disease 2019 (COVID-19) vaccines. It was demonstrated that the higher the predicted score from the DDR model, the more progress in the clinical phase of the COVID-19 vaccine development. Although our approach has limitations that should be improved, it makes scientific as well as industrial contributions in that the DDR model can support rational decision-making prior to initiating drug development by considering not only technical aspects but also company-related variables.

1. Introduction

New drug development is a high-risk, high-return business that guarantees enormous profits upon success, but the success rate is extremely low (Munos, 2009; Taylor, 2016). It requires extensive research and development (R&D) periods including clinical trials and immense investment costs to develop just a single drug. For these reasons, pharmaceutical companies have eagerly sought various strategies and tactics to increase the possibility of success in their drug development projects (Schuhmacher, Gassmann, & Hinder, 2016). One of these solutions can be to take advantage of big data and artificial intelligence (AI), which have recently undergone exponential growth and expansion (Chang, Chang, & Huang, 2020; Henstock, 2019; Tseng, Tran, Ha, Bui, & Lim, 2021).

The pharmaceutical industry has been vigorously adopting data mining, machine learning, and AI technologies to reduce the time and cost required for drug discovery and development, mainly led by multinational large pharmaceutical companies (Jiménez-Luna, Grisoni, & Schneider, 2020; Reda, Kaufmann, & Delahaye-Duriez, 2020). The value of AI in the drug discovery market is expected to grow rapidly at a compound annual growth rate of 40.8%, increasing from 260 million US dollars (USD) globally in 2019 to reach 1.43 billion USD in 2024 (MarketsandMarkets, 2019). The areas in which AI technologies can be applied in drug discovery and development include target identification and validation, small-molecule design and optimization, prediction of biomarkers, and computational pathology (Vamathevan et al., 2019). Specifically, most of the applications have been focused on optimizing and enhancing efficiency in the drug discovery process (Chan, Shan, Dahoun, Vogel, & Yuan, 2019; Freedman & Reardon, 2019; Schneider et al., 2020) that can be segmented into sub-categories of compounds, genomics, targets, and antibodies (Schuhmacher, Gatto, Hinder, Kuss, & Gassmann, 2020).

However, in addition to the experimental drug discovery process, it is necessary to smartly leverage AI technologies in the planning stage of drug development in terms of business and management. Pharmaceutical companies wonder what kind of drugs they should aim for their next development project, considering their company’s situation from managerial and market perspectives. In most cases, the decisions on the
drugs aimed at development have been made qualitatively based on the views of the company’s board of directors or the technologies possessed by the companies (Schuhmacher, Gassmann, Hinder, & Kuss, 2021). That is, decision-making on drug development has hitherto depended more on human judgment than formal analytical or data-based methodologies (Jekunen, 2014). Therefore, there is a great need for an effective model or solution that can increase their success rate of drug development.

In this regard, this study arose from the following research questions: “Can AI be utilized for decision-making on drug development plans? Specifically, can data-based machine learning techniques suggest developable drugs depending on the business situation of each pharmaceutical company?” In this context, the objective of this study is to develop a decision-support model to recommend the most developable drug groups in a manner customized for individual pharmaceutical companies. The current technologies recommend new drug candidates that are expected to be effective in particular diseases based on technological factors (Jeon et al., 2014; Wang, Gu, Wei, Cao, & Liu, 2015; Zhang, Wang, & Hu, 2019) regardless of the actor to perform the development. That is, the previously reported methods of recommending new drug candidates were difficult to be tailored for a specific company. When generating recommendations, they did not take into account which companies would be most effective at developing a specific drug and which drugs have the highest probability of success for a specific company. However, the possibility of development success can be greatly influenced by factors such as the company’s drug development experiences, technical know-how, product portfolio, financial status, and the market environment of the drugs (Arora, Gambardella, Magazzini, & Pammolli, 2009; Blau, Pekny, Varma, & Bunch, 2004; Ringel, Tollman, Hersch, & Schulze, 2013). In this study, we tried to address these issues, which were insufficiently considered in prior research, by proposing a decision support model that recommends new drugs that is most suitable for a company (or developer), considering not only technological aspects but also management/business aspects. We have adopted the main conception of the recommender systems with the objective of devising a company-tailored drug development recommendation (DDR) model that considers each company’s situational conditions.

Specifically, we developed a new model that recommends and/or predicts drug groups with high probability of success in development. The proposed model is based on the hypothesis that “users’” information in recommender systems can correspond to “pharmaceutical companies’” and “purchased/rated items’” information can be matched with “drugs developed successfully.” Among the techniques employed in recommender systems, we applied three types of approaches, namely association rule learning, collaborative filtering (CF), and content-based filtering (CBF), which are the most widely and successfully utilized algorithms. The results obtained through each approach were then combined to create a hybrid model, named the DDR model, in order to compromise the limitations of the individual approaches and improve the quality of recommendations/predictions. The DDR model recommends and/or predict drug groups suitable for development in a highly enterprise-customized (i.e., personalized) manner. This study added scientific value to the literature by investigating the features which were not fully considered in previous studies and identifying several important characteristics related to companies and drugs that can predict drug development success. Also, it presents a reliable method to improve the quality of recommendation by combining three different algorithms for the purpose of use in drug development planning. In addition, it offers industrial contributions that it can effectively be utilized in the product development process to assist decision-making or prioritization about which drug groups to develop, before undertaking the drug development process. It is also feasible to predict which companies are most likely to succeed in the development for a specific drug group, among the companies currently engaged in its development, based on the DDR model. The validity and practical applicability of the DDR model was successfully demonstrated by accurately predicting the companies with high potential of success in the development of COVID-19 vaccines, one of the drugs most urgently in need of development.

2. Theoretical backgrounds

2.1. AI in pharmaceutical portfolio management

AI technologies have been actively introduced in drug discovery and development since the last few decades. The current state of AI in the pharmaceutical industry is considered to be at an early mature level in the technology life cycle (Schuhmacher et al., 2020). As one of the AI applications, it is utilized not only in the experimental process of drug discovery and development, but also in drug portfolio management. Portfolio management is known as a dynamic decision process that facilitates the evaluation, selection, and prioritization of new projects, and the acceleration, discontinuation, and deprioritization of existing projects (Cooper, Edgett, & Kleinschmidt, 1999). Pharmaceutical companies have utilized it as one of the most important strategies to increase R&D success rates by reducing project risks (Schuhmacher et al., 2016). Ding et al. (Ding & Ellisberg, 2002) and Jekunen (Jekunen, 2014) identified that the factors mainly considered to construct an optimal pipeline portfolio include the cost of development, likelihood of surviving, expected profitability, competitive situation, market size, and novelty of the drug. In order to make decisions about drug development, pharmaceutical companies have relied heavily on human judgement and prior product development experiences in the past (Betz, 2011; Krishnan & Ulrich, 2001). To address the limitations of these qualitative decisions, quantitative data-driven methods, such as scoring, surveys, discounted cash flow (DCF), real option valuation (ROV), Monte Carlo/descrete event simulations, patent evaluation methods have been applied to lower risks and maximize returns (Betz, 2011; Blau et al., 2004; Jekunen, 2014). More recently, AI technologies such as machine learning or deep learning are expected to rapidly replace much of project and portfolio management in pharmaceutical R&D (Schuhmacher et al., 2021).

Specifically, there have been several attempts to predict the success probability and likelihood of approval (LoA) of clinical trials based on diverse factors affecting the success of the clinical trials, and to utilize it for pharmaceutical portfolio decision-making. DiMasi et al. (DiMasi et al., 2015) developed a predictive model using statistical methods (associations and logistic regressions) and machine learning techniques (random forest and classification and regression trees (CART)) to predict regulatory marketing approval after phase II in the case of cancer drugs. They proposed a scoring method that combines the results of these techniques, called the Approved New Drug Index (ANDI) metric. They demonstrated that the predictive performance of the ANDI metric obtained using only four factors (number of patients in pivotal phase II, number of patients treated worldwide, phase II duration, and activity (response rates)) was sufficiently high for predicting the regulatory marketing approval of cancer drugs. In addition, Lo et al. (Lo, Siah, & Wong, 2019) developed a model to predict drug approvals based on machine learning using the data on drug development and clinical trials from 2003 to 2015. In particular, they applied statistical imputation methods to missing values to utilize the available data as efficiently as possible. The model built by applying this approach was superior in performance compared to complete case analysis, which generally applied in most previous studies. The important features found to predict drug approvals in this model were trial outcomes, trial status, and trial accrual rates. They asserted that the results obtained from this study offer useful insights into the outcome of drug development because many of these variables that had not been considered in prior studies.

Feijoo et al. (Feijoo, Palopoli, Bernstein, Siddiqui, & Albright, 2020) also developed a model to predict the phase transition and LoA of clinical trials using supervised machine learning (SML) and Natural
Language Processing (NLP) algorithms. The authors used the NLP algorithm to extract an indicator measuring the complexity of eligibility criteria from the text data of ClinicalTrials.gov, which had not been explored in previous studies. Including this predictor, the authors constructed a SML model which can predict the clinical phase success with an accuracy of 80%. This study found that the eligibility criteria complexity and the number of end points were key predictors. They suggested that the SML model can be used to obtain insightful information for clinical protocol design or operational feature management (from the clinical trial practitioners’ perspective) and portfolio assessment and effective decision-making (from the entrepreneurs’ perspective).

Since most of the existing prediction models used the characteristics related to the conditions or results of clinical trials as important variables for prediction, the scope of application of such models may be limited to drugs after the initiation of clinical trials. However, many companies need to forecast which drug will have a high probability of success before starting drug development, that is, in the business planning stage, for effective drug portfolio management and decision-making on investment prioritization. In this context, the DOR model proposed in this study has the advantage of being able to predict in advance the possibility of drug development success using company profiles and a few characteristics of a drug even before the start of clinical trials. In addition, by combining the algorithms used in the recommender systems, which have been widely used for product purchase recommendation but not utilized for drug development planning, it can provide not only the possibility of clinical trial success but a recommendation ranking for specific drug classes (from among 77 drug classes covering all disease areas), tailored to each pharmaceutical company. A more detailed discussion of the recommender systems is continued in the next section.

2.2. Recommender systems

Recommender systems, which are one of the most successful applications of AI, suggest preferable items to customers based on user-item interaction information (Adomavicius & Tuzhilin, 2005). Its practical usefulness and effectiveness have been substantially proven by large-scale business applications in industries such as YouTube, Amazon, and Netflix. The recommender systems can generally be categorized into association rule mining methods and information filtering methods. The association rule mining is a traditional data mining technique to find association rules between a set of co-purchased products (Sarwar, Karypis, Konstan, & Riedl, 2000a). Apriori, tree projection, and direct hashing and pruning (DHP) algorithms have been widely used to find association rules from various kinds of transaction databases (Huang, Chen, Wang, & Chen, 2000).

The information filtering methods are classified into CBF approach and CF approach (Adomavicius & Tuzhilin, 2005). The CBF approach is a method of analyzing the contents of the items and/or user profiles and recommending items that are similar to those of previously preferred by a particular user (Lu, Wu, Mao, Wang, & Zhang, 2015). In order to analyze the content of the items, many CBF approaches focus on extracting a set of features from textual information utilizing text mining techniques such as term frequency-inverse document frequency (TF-IDF) measure or semantic analysis (Adomavicius & Tuzhilin, 2005). The user profile is used to predict the users’ preferences even if their past purchase/rating history or other users’ item evaluation scores do not exist (Isinkaye, Polajimi, & Ojokoh, 2015). In the CBF approaches, two techniques have been used to generate recommendations. One technique utilizes traditional information retrieval methods, such as cosine similarity measure, to heuristically generate recommendations. The other technique utilizes machine learning methods to build a model that learns users’ interests from the training dataset of users followed by recommendations generation (Pazzani & Billsus, 1997).

The CF approach predicts users’ preferences on items based on user-to-user or item-to-item similarity from the basic assumption that customers with similar preferences for a particular item will have similar preferences for other items. In other words, the CF approach is a method to select other users with the similar preferences as target customers and recommend their preferred items to the target customers. Since Goldberg et al. (Goldberg, Nichols, Oki, & Terry, 1992) introduced this concept for the first time, there have been numerous applications over the past decades in academia and industries (Bobadilla, Ortega, Hernandez, & Gutierrez, 2013). The algorithms used for the CF can also be divided into memory-based and model-based techniques. In the memory-based technique, recommendations are generated through a heuristic way by performing similarity measurements and preference predictions. In the model-based technique, machine learning algorithms such as classification, clustering, and dimension reduction approaches are utilized to build a recommendation model (Hofmann, 2004; Lu et al., 2015).

However, the CBF approach has problems with limited content analyses and overspecialization (Adomavicius & Tuzhilin, 2005). The CF approach also has challenges with cold-start, sparsity, and scalability of data (Isinkaye et al., 2015). In order to address these issues, many hybrid recommender systems have been developed that integrate different approaches to produce high quality recommendations (Ai Mamun Rashid, Karypis, & Riedl, 2006; Barragans-Martinez et al., 2010; Burke, 2002; McNee, Riedl, & Konstan, 2006). When constructing a hybrid recommender system, a variety of combination techniques are used, such as weighted averaging of each algorithm’s recommendation scores (Burke, 2002), combining recommendation rankings for each algorithm (Pazzani, 1999), applying the results of the CF approach into the CF approach or vice versa (Adomavicius & Tuzhilin, 2005; Paulson & Tzanavari, 2003), and selecting the most appropriate recommendation engine in real-time by recognizing the current situation from several already learned recommendation engines (Ducheneau et al., 2009). In this study, we have developed a hybrid model to recommend and/or predict drug groups suitable for development by individual pharmaceutical companies, by adapting the principles of recommender system and combining the approaches of association rule learning, CBF, and CF.

3. Methods

3.1. Data collection and refinement

The data on the drugs marketed over the last three years by pharmaceutical companies was obtained from IQVIA™ Pipeline Intelligence. In order to reflect the latest product development trends, the data collection period was constrained to the last three years (from May 2017 to April 2020). Information such as product name, drug class code, indication, mode of administration (MoA), and developing company was included in the collected data. The main training dataset in the form of the company-drug counting matrix, consisting of the number of marketed drugs by pharmaceutical company and drug class, was prepared from this raw data for all approaches in this study. In addition, the data on the market sizes of the drugs was obtained from IQVIA™ therapeutic class profiles and calculated based on the second level Anatomical Classification (AC) of drug class (EphMRA/Intellus Classification Committee, 2019).

The data regarding the pharmaceutical companies for which the drug information was collected as described above were obtained from Standard & Poors Capital IQ. Basic information about the companies such as country/region of incorporation, year founded, and number of employees and financial information such as market capitalization, total revenue, and R&D expense (a total of 27 variables) were acquired and utilized for machine learning in the CBF approach. Given that the information on the drugs covered the last three years, we chose to use company data for 2017 to reflect the time-lag effect.
3.2. Association rule learning

Association rule learning was performed based on the main dataset on the number of marketed drugs by company and drug class. Association rule learning can find interesting patterns or rules in the relationship between item sets in large transactional data (Agrawal, Imielinski, & Swami, 1993; Ozseyhan, Badur, & Darcan, 2012). Our experiment employed the Apriori algorithm, which is the most widely utilized algorithm for investigating association rules (Agarwal & Srikant, 1994). In order to discover significant rules latent in the dataset, the three metrics of support, confidence, and lift of the rules were measured by the following formula (Hornik, Grün, & Hähls, 2005; Mining, 2006) and their minimum thresholds were applied:

\[
support(X) = \frac{frequency(X)}{N} \tag{1}
\]

\[
confidence(X\rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)} \tag{2}
\]

\[
lift(X\rightarrow Y) = \frac{supp(X \cup Y)}{supp(X) \times supp(Y)} \tag{3}
\]

The original definitions were modified as follows: \(N\) denotes the total number of pharmaceutical companies (drug developers) in the dataset and \(frequency(X)\) denotes the number of companies that include drug class \(X\) in their product portfolio. Likewise, \(support(X \cup Y)\) denotes the proportion of companies containing drug class \(X\) and drug class \(Y\) simultaneously in their product portfolio. As a result of applying minimum threshold criteria of 0.03 for support and 0.8 for confidence, 1,834 association rules were discovered for drug development.

3.3. CF modeling

The main dataset on the number of marketed drugs by company and drug class was adjusted for the CF approach. Any number of marketed drugs above five was convergent to five to compensate for problems that may be caused by a large deviation of the number of the marketed drugs by pharmaceutical company. In addition, the data on companies possessing less than four drug classes were filtered out to ensure there would not be any companies that do not have items to test when evaluating the CF model.

The CF modeling was performed using this modified dataset by employing different algorithms of user-based collaborative filtering (UBCF), item-based collaborative filtering (IBCF), singular value decomposition (SVD), and funk singular value decomposition (SVDF). In the UBCF and IBCF algorithms, cosine similarity was used when calculating the similarity between users or items. The performance of each algorithm was compared to select the best algorithm for the CF model. First, the accuracy of the algorithms was assessed by root mean square errors (RMSE) and mean absolute error (MAE) between the actual number of drug developments and the predicted number of drug developments by the CF model. Secondly, we converted the subject into a binary classification problem of development (or not) and evaluated the accuracy of the algorithms using the receiver operating characteristic (ROC) curve.

3.4. CBF modeling

For the CBF model generation, feature selection was preceded in order to identify an optimal set of predictive variables and avoid overfitting of the data. The filter method that examines the statistical validity of the features and a wrapper method, which selects variables by repeating the task of performing modeling using a part of the variables and checking the results, were applied for the feature selection.

In the filter method, the variance of each feature was checked and the features with variances close to zero were removed because this indicates they rarely have different observations. In our variables on the dataset, the “perrectum” and “sublingual” modes of drug administration and the “depreciation and amortization” and the “selling and marketing expense” of a company showed variances close to zero. Therefore, these four variables were removed. Next, the correlations between the variables were examined to remove the variables with high correlation coefficients. This is because when there are highly correlated variables in the variables set, the performance of the machine learning model may deteriorate or the model may become unstable (Kühn, 2008). Among the variables with a correlation coefficient of 0.6 or higher, the following four variables were excluded, except for total revenue and R&D expense, in consideration of representativeness: total enterprise value, gross profit, EBITA, and number of total investments and subsidiaries of a company.

In the wrapper method, the Boruta algorithm, which is a feature ranking and selection algorithm, was applied. The concept of the Boruta algorithm is to remove variables that do not affect a model more significantly than the shadow features obtained by shuffling the values of the original attribute across objects (Kursa & Rudnicki, 2010). “Company status” was a variable that was found to be unimportant through the Boruta algorithm. Therefore, after removing this variable last, a total of 31 variables were applied for the CBF modeling (Table 2 and Table S1 and S2 in Supplementary material).

The main dataset of company-drug counting matrix was expanded by adding company profiles and drug features chosen from the feature selection process. The missing values in the company profiles were imputed by the median values according to the classification of country/region of incorporation. The number of marketed drugs by class code and company was converted into a binary class (0 and 1) and used as a dependent variable indicating the success of drug development for the supervised learning classification in the CBF model. Since the deviation between the frequency of two classes was large, they were balanced by the up-sampling method (Van Hulse, Khoshgoftaar, & Napolitano, 2007).

The classification algorithms of decision tree, random forest, support vector machine (SVM), and k-nearest neighbors (kNN) were applied for the CBF modeling. The hyperparameters for each algorithm were optimized using the caret package in R. The CBF model performance was evaluated by 5-fold cross validation and by using the metrics of accuracy, sensitivity, specificity, and area under curve (AUC). The importance of the variables was investigated by mean decrease in accuracy and the Gini index for the features from the random forest-based CBF model.

3.5. DDR model construction

The DDR model was constructed by incorporating the results of association rule learning, CF, and CBF approaches. It was designed to ensure that recommendation scores were derived for all drug classes of all pharmaceutical companies used for training in this model.

Specifically, regarding association rule learning, the drug classes in the consequents (right hand side; RHS) of the generated rules are recommended when the drug classes in a particular company’s product portfolio are included in the antecedents (left hand side; LHS) of the rules. The lift values of the generated rules are used as a recommendation score. As for the CF approach, the results from utilizing the SVD algorithm were applied to the DDR model. The predicted counting values were obtained for the drug classes that had not yet been developed by each company, and the actual values were returned for drug classes that had already been developed. For the CBF approach, the results of employing random forest algorithm were applied to the DDR model. The probability values for the success of development obtained for each drug class for each company were used as a recommendation score.

These different recommendation scores were incorporated in the DDR model using a weighted linear combination method. The scores
obtained from each approach were normalized since they used different scales when generating recommendations. The total score (S) was calculated using the following formula:

$$S_{ij} = \sum_{i \in R \cap C \in D} w_i \left( \frac{r_{ij,k} - r_{ij,\text{min}}}{r_{ij,\text{max}} - r_{ij,\text{min}}} \right)$$  \hspace{1cm} (4)$$

Where $w_i$ is a weight proportion of the recommendation approach $i$ such that $\sum w_i = 1$, $r_{ij,k}$ is the recommendation score of company $j$’s drug class $k$ in the recommendation approach $i$, $r_{ij,\text{max}}$ and $r_{ij,\text{min}}$ are the maximum and minimum values of the recommendation score in the recommendation approach $i$, respectively, $R$ is the set of all recommendation approaches (association rule learning, CF and CBF), $C$ is the set of all pharmaceutical companies, and $D$ is the set of all drug classes. The priority of the drug recommendation is derived based on $S$ within a particular company.

3.6. Implementation

The DDR model was developed under R version 4.0.2. and relies on the following packages: arules, arulesViz, Recommenderlab, caret, Boruta, and randomForest. The custom code created in this study is available upon reasonable request.

4. Results

4.1. A framework of the DDR model and summary of input data

An overview of the DDR model is schematically illustrated in Fig. 1. First, the data on successfully developed (i.e., marketed) drugs classified by company was collected as the main dataset for learning in the DDR model. In addition, the data on company profiles and drug features are also collected and learned. In the DDR model, associate rule learning, CF, and CBF approaches, which are some of the most fundamental and widely utilized algorithms for recommender systems, are adopted and modeled to generate recommendations and predictions. The recommendation scores obtained through the three approaches are incorporated as a hybrid model in order to compensate for the limitations of individual algorithms and improve the quality of the recommendations. Finally, the DDR model outputs the recommendation of the drug classes suitable for development for each company and/or the prediction of the companies with high probability of success for each drug class.

As for input data entered in the DDR model, the main dataset consists of information about the number of drugs marketed during the last three years classified by pharmaceutical company and by drug class (Fig. 1 and Table 1). The classes of drugs were in conformity with the AC of pharmaceutical products established by the European Pharmaceutical Market Research Association (EphMRA). In the AC system, drugs are classified according to anatomical site of action, their indications, therapeutic use, composition, and mode of action and hierarchically organized into four levels (EphMRA, 2020). The first level indicates anatomical main group (e.g., C: cardiovascular system) and the second level identifies therapeutic main group (e.g., C10: lipid-regulating/anti-atheroma preparations). For all the experiments in this study, we constructed the dataset categorized in the second level of the AC code.

The number of pharmaceutical companies and drug classes included in the dataset were 957 and 77, respectively (Table 1). It was found that the number of drug classes marketed by pharmaceutical companies ranged from one to forty-two, with an average of 3.66 classes. In addition, the other dataset consisting of company profiles and drug features was also entered in the DDR model, particularly for CBF in the model (Fig. 1 and Table 2). The drugs included in the dataset had single or multiple administration modes and there was originally a total of 12 administration types. Through the feature selection process, the “per-rectum” and “sublingual” types were found to be non-informative, so they were not used as predictor variables. The market size of the drug classes was found to have a value between 69 million USD and 98,331 million USD. A summary of 20 characteristics on the pharmaceutical companies with marketed drugs is listed in detail in Table 2 and Table S1 and S2 in Supplementary material.

4.2. Discovering association rules on drug development

The components of ‘transactions’ and ‘items’ in the general association rule mining approach were substituted as ‘drug developments’ of pharmaceutical companies and ‘drug classes’, respectively in the DDR model. It was observed that the highest number of developed drugs over

Table 1

| Number of pharmaceutical companies | Number of drug classes | Number of drug classes by company* |
|------------------------------------|-----------------------|-----------------------------------|
| Mean (SD)                          | Median (IQR)          | Min. Max.                         |
| 957                                | 77                    | 3.655 (5.121)                     |
| 2 (3)                              | 1                     | 42                                |

* Multiple counting was allowed if a single drug corresponds to multiple drug class codes and if several companies are involved in the development of a single drug.
the last three years were in the V7 class (all other non-therapeutic products, 8.15%), L1 class (antineoplastics, 6.15%), N7 class (other CNS (central nervous system) drugs, 4.28%), L4 class (immunosuppressants, 3.77%), and M1 class (anti-inflammatory and anti-rheumatic products, 3.14%).

For 613 pharmaceutical companies, excluding the companies with only one marketed drug, 1,834 association rules were generated on drug development, by applying 0.03 for the minimum support threshold and 0.8 for the minimum confidence threshold (Table 3 and Fig. 2). Out of 1,834 rules, the largest share were rules related to the L4 class (immunosuppressants), followed by the L1 class (antineoplastics) and the M1 class (anti-inflammatory and anti-rheumatic products). It can be interpreted that since these are drug groups for cancers or infectious diseases that are major diseases, recording high incidence and prevalence, pharmaceutical companies have developed many drugs related to these diseases, and accordingly, many rules for them have been discovered.

The mean values of support, confidence, and lift were 0.036, 0.906, and 5.002, respectively, suggesting that strong association rules were identified (Table 3). The rules were derived from the combinations of two to seven drug classes, as shown in the length of rules in Table 3. The top twenty association rules based on the lift value are shown in Fig. 2b. The rules where the L2 (cytostatic hormone therapy), N5 (psycholeptics), N6 (psychoanaleptics excluding anti-obesity preparations), and R3 (anti-asthma and COPD (chronic obstructive pulmonary disease) products) classes came out as the consequents (RHS) were found to have high lift values. In particular, the lift value of the rules was highest for [H1 class (pituitary and hypothalamic hormones)] → [L2 class (cytostatic hormone therapy)] (i.e., if the H1 class is developed, then the L2 class is also developed.). It is assumed to be due to high similarity of both classes as hormonal drugs.

Recommendations are generated based on the discovered association rules in the following manner: if the drug development portfolio of a particular company is subordinate to the antecedents (LHS) of the generated rules, the drug classes in the consequents (RHS) of the rules are recommended to the company. The prioritization of the recommendations follows the order of the lift values of the corresponding rules. For 26.3% of the total of 957 companies, the recommendations were generated because their drug portfolios were included in the antecedents of 1,834 discovered association rules.

### 4.3. CF modeling and its performance

CF modeling was conducted based on the 'company-drug counting matrix' consisting of the number of developed (marketed) drugs by pharmaceutical company and by drug class, which replaced the 'user-item rating matrix' commonly used in CF approaches. We trained the data from the companies with at least four of the developed drug classes for the CF modeling in this study (They accounted for 26.2% of the main dataset). The histogram showing the frequency of the number of developed drugs per drug class of a company is presented in Fig. S1 in Supplementary material.

We evaluated the performance of the CF model by 5-fold cross validation with 80% of the training data and 20% of the test data. In the evaluation process, the performance of the CF model utilizing UBCF, IBCF, SVD, and SVDF algorithms were compared with that of random-manner recommendation.

In the case of generating ten items recommendation, it was observed...
that the values of the evaluation metrics, RMSE and MAE, of all algorithms except for IBCF were lower than those of random recommendation (Table 4). The SVDF algorithm showed the lowest value for RMSE, and the UBCF algorithm showed the lowest value for MAE. In addition, the ROC curves for each algorithm were drawn and compared to select the best algorithm, by converting the question to a binary classification problem (Fig. 3). That is, we considered the problem as having the purpose of classifying the drugs by recommendation/non-recommendation and development/non-development. In the case of generating 1 to 20 recommendations, the SVDF algorithm showed the largest value of the AUC, compared to the AUCs obtained from all algorithms. Therefore, the SVDF was chosen as a representative algorithm of the CF approach to be applied to final DDR model.

4.4. CBF modeling and its performance

A supervised machine learning algorithm was applied to predict the possibility of development success for each drug class by company, based on a total of 31 variables including drug development portfolios, company profiles, and drug features (refer to Table 2 and Table S1 and S2 in Supplementary material). The CBF modeling was performed by applying various types of classification algorithms, namely decision tree, random forest, SVM, and kNN, to select the optimal algorithm with the strongest performance. The prediction scores on development success were computed by the CBF model for all drug classes that had not yet been developed by the companies included in the input dataset, and the recommendations were generated in the order of high prediction scores.

The evaluation of the CBF model applying each algorithm was performed by 5-fold cross validation and the performances for the test dataset were compared using the metrics of accuracy, sensitivity, specificity, and AUC. The random forest algorithm showed the best performance among them when considering the evaluation metrics comprehensively (Table 5), and therefore it was adopted as a representative algorithm of the CBF approach for application to the final DDR model.

In the ROC curve of the random forest classifier, the AUC was 0.74, indicating it to be capable of properly classifying the development success of drug class (Fig. 4a).

In addition, the importance of predictor variables was investigated by finding the mean decrease in accuracy and node impurity (Gini index) from the random forest-based CBF model. It was found that the total number of marketed drugs of a company, whether the drug administration modes are oral and injection, and the market size of the drug class are the most important predictors for drug development success (Fig. 4b).

4.5. DDR model building and its empirical validation

The DDR model was established by combining three individual approaches. The recommendation scores derived from each approach were

| Algorithm | RMSE  | MAE  |
|-----------|-------|------|
| UBCF      | 0.963 | 0.739|
| IBCF      | 1.210 | 0.928|
| SVD       | 0.962 | 0.777|
| SVDF      | 0.937 | 0.781|
| Random    | 1.204 | 0.918|

Fig. 2. (a) Confidence-support plot on the generated 1,834 association rules on drug development. The confidence and support values of each generated rule are shown as a scatter plot. (b) A grouped matrix with antecedent groups (LHS) as columns and consequents (RHS) as rows on the top 20 association rules based on the lift value of the generated rules. The lift value decreases from top to down and from left to right (i.e., the color of the bubble varies from red to grey in descending order.).

Fig. 3. Comparison of the ROC curve of the algorithms applied in the CF approach. The number from 1 to 20 indicates the number of recommended items (drug classes).
normalized and summed with weighted combination ratios. Fig. 5 shows a part of the results of recommending drug classes for development tailored to an individual company, based on the DDR model incorporating each approach’s scores with an equal weight combination ratio. The x-axis and y-axis in Fig. 5 indicate drug class and pharmaceutical company, respectively. (The company names were filled anonymously.) It was observed that the priorities of the drug classes recommended for development were obtained differently for each company, as shown by the color variation in the heatmap. Specifically, in the cases of companies with a drug development portfolio that satisfied the discovered association rules and companies that obtained recommendation scores over the last three years, the developing company’s success rate of 12.9% over the past decade (IQVIA Institute, 2020). In this regard, a rational and data-driven decision-support system is needed to help the companies determine the drugs with high potential for success as their next development goals, tailored to their individual circumstances, in order to reduce risks and maximize profits from the development.

In this study, we developed a decision-support system, named the DDR model, that can effectively guide pharmaceutical companies’ new drug development planning by leveraging machine learning. Machine learning algorithms such as supervised/unsupervised learning methods have been widely applied in various fields, but they have rarely been utilized in business decisions on drug development, notwithstanding a large amount of data being produced every day regarding drug development planning. The DDR model was constructed by incorporating association rule learning, SVDF-based CF, and the random forest-based CBF model, taking advantage of machine learning and low success rates in new drug development (Jung, Hwang, & Yoo, 2020; Khanna, 2012; Thakor et al., 2017). Consequently, pharmaceutical companies are facing serious challenges to their business models (Paul et al., 2010). The composite success rate for drugs, which indicates the success of products from clinical phase 1 to the regulatory decision for marketing, has declined in recent years, with an average success rate of 12.9% over the past decade (IQVIA Institute, 2020). In this regard, a rational and data-driven decision-support system is needed to help the companies determine the drugs with high potential for success as their next development goals, tailored to their individual circumstances, in order to reduce risks and maximize profits from the development.

Despite astronomical and unparalleled R&D investment, the pharmaceutical industry has continued to struggle with low R&D productivity and low success rates in new drug development (Jung, Hwang, & Yoo, 2020; Khanna, 2012; Thakor et al., 2017). Consequently, pharmaceutical companies are facing serious challenges to their business models (Paul et al., 2010). The composite success rate for drugs, which indicates the success of products from clinical phase 1 to the regulatory decision for marketing, has declined in recent years, with an average success rate of 12.9% over the past decade (IQVIA Institute, 2020). In this regard, a rational and data-driven decision-support system is needed to help the companies determine the drugs with high potential for success as their next development goals, tailored to their individual circumstances, in order to reduce risks and maximize profits from the development.

This empirical analysis demonstrates the validity of the DDR model.

### Table 5
Comparison of the evaluation metrics of accuracy, sensitivity, specificity, and AUC of the algorithms applied in the CBF approach.

| Classifier   | Accuracy | Sensitivity | Specificity | AUC  |
|--------------|----------|-------------|-------------|------|
| Decision tree| 0.653    | 0.796       | 0.645       | 0.721|
| Random forest| 0.775    | 0.708       | 0.799       | 0.744|
| SVM          | 0.826    | 0.625       | 0.838       | 0.731|
| kNN          | 0.742    | 0.617       | 0.749       | 0.683|

![Fig. 4. (a) The ROC curve for the random forest-based CBF model. (b) The variable importance on the success of drug development obtained from the random forest-based CBF model.](image-url)
factorization technique designed to reduce the dimension of data by decomposing the data matrix into smaller ones (Sarwar, Karypis, Konstan, & Riedl, 2000b). It is known that the SVD method improves the performance of the CF approach by overcoming the sparsity problem of the user-item rating matrix (Sarwar et al., 2000a, 2000b). We employed a SVDF method popularized by Simon Funk, which decomposes a matrix by stochastic gradient descent optimization to minimize errors in the known values (Koren, Bell, & Volinsky, 2009).

As for the CBF approach, its performance was superior even as a single model and there was no significant difference in performance depending on the tested classification algorithms (Table 5). These results prove that our proposal of adapting a principle of recommender system to drug development planning is reliable and the model has been built successfully. In addition, the CBF approach has an advantage over association rule learning or the CF approach because it can generate prediction scores followed by recommendations for all drug classes in any company, regardless of past drug development experiences or portfolio composition of drug classes.

We also identified from the CBF model that the total number of drugs developed by the company and the company’s financial status, including factors such as total liabilities, total equity, and revenues are the major factors affecting the success of drug development. The duration of business operation and the number of employees were also important variables. This is consistent with the findings in previous studies in other industries that a company’s product development experiences and/or the size and history of a company significantly influence new product development success (Ernst, 2002; Murphy & Kumar, 1996). More specifically, in the previous study on the prediction of the clinical approval of drugs, it was found that firms with larger sales had a higher rate of regulatory approval of oncology compounds than firms with lower sales (DiMasi et al., 2015). However, the financial conditions of a company have been rarely explored in detail, so by including these variables in this study, the factors in the financial management aspects of a company that can predict the success of drug development were identified.

In terms of drug features, the type of administration (oral, injection, and topical) was found to be an important factor and the market size of the drug class was also an important explanatory variable. It is in line with the finding of DiMasi et al. that oral and injectable modes of administration were associated with the clinical approval of new cancer drugs (DiMasi et al., 2015). Orally administered drugs (32.7%) showed a higher probability of clinical trial success than the injectable drugs (10.2%) in the authors’ study. Regarding the market size, previous studies focusing on other industries have already revealed that the market size of products is one of the key factors in R&D success (Astebro, 2003; Balachandra & Friar, 1997). We also found in this study that market size is a major factor affecting the success of new drug development in the pharmaceutical industry. The CBF model showed a comparable performance with the existing models, even though it did not use the conditions for clinical trials or disease characteristics as predictor variables.

It may be concerned that the results of the CBF model only made an obvious prediction that companies which completed a lot of drug developments in the past would perform well in the future. However, we...
found other significant predictors, as well as the number of marketed drugs, regarding drug features and company profiles. For example, in order to increase the success rate of drug development, companies can strategically choose the drug’s administration type to be an oral or injection type, and intentionally focus on the development of the drugs with a large market size. It can also be inferred from the results that managing a company’s total debt will be more beneficial to its drug development success. Importantly, the DDR model does not only predict the success probability of a drug by using these variables (in the CBF model), but it also finds the rules for drug development from the drug development portfolios of pharmaceutical companies across the world (in the association rule mining). Also, by calculating the similarity in drug development of pharmaceutical companies, it can provide insights into which of the 77 drug classes would be relatively more advantageous to develop for a specific company (in the CF model).

Since the individual approaches we employed had their pros and cons, they were combined as a hybrid model to reinforce their strengths and compensate for their weaknesses. Three approaches in the DDR model were combined by averaging the normalized recommendation scores of each approach with an identical weight ratio. However, different weights can be applied depending on the business situation of the target company. For instance, for companies that have a relative abundance of experience in drug development, a higher weight ratio is given to the CF approach in the DDR model, and for companies with cold start problems due to the lack of a drug development history, a higher weight ratio is applied to the CBF approach to generate recommendations. In the validation stage of the DDR model, we used the prediction scores obtained solely by the CBF approach to provide an accurate value of each company’s success probability for specific drug class development.

### 6. Implications

Considering the high cost and risk of new drug development, a more prudent utilization of the recommendation model is required, compared to product purchases or preferable contents recommendations. In this study, the significant rules and reliable model performances were achieved from each algorithm (association rule learning, CF, and CBF approach) in the DDR model. It implies that the principles of recommendation model can effectively be applied to establish a decision-support model for drug development plans. In addition, a hybrid model combining three algorithms was presented in this study to improve the quality of recommendations. In this regard, this study makes a theoretical contribution in that it suggests a new methodology that can be used in the purpose of product development planning. Furthermore, this study identified the predictors related to financial management of a company that can predict the success of new drug development, which have been hardly addressed in prior research.

From practical (managerial) perspectives, pharmaceutical companies can benefit from the DDR model in making rational decisions about which drug (classes) to develop for the purpose of business planning or portfolio management. It may contribute to reduce the time and cost invested in new drug development of the companies. Furthermore, it provides implications for the government agencies in charge of public health that a data-driven approach as adopted in our model can be used to predict the success probability of drugs under R&D in order to preemptively secure drugs that urgently needed for public health, especially in an emergency situation. Even drugs that have not yet been developed can be purchased in advance by predicting the likelihood of success for the safety of the public. In the wake of COVID-19, the emergence of another new strain of virus is widely predicted, and as the occurrence cycle of such viruses is expected to be shortened, it is essential for the public sector to prepare for the threat of another pandemic.

### 7. Conclusions

To the best of our knowledge, this study devised a novel recommendation/prediction model that can help overcome the challenge that the success rate of drug development has not improved for quite a long time. The DDR model is expected to contribute to cutting down the time and cost required for new drug development in the private sector by suggesting the most developable drug groups, highly customized for individual companies’ circumstances. Furthermore, using the DDR model in the public sector, national investment and support efforts can be strategically strengthened for companies developing drugs that can improve public health, and systems can be put in place to preemptively respond to the sudden emergence of new infectious diseases such as SARS, MERS, and COVID-19, because the DDR model can predict which companies are more likely to succeed in developing particular drugs in urgent need. In this context, the DDR model can be an attractive and essential tool utilized as a decision-support system adaptively in both the private and public sectors.

Since this is just an initial step to showcase the feasibility of our proposed concept, the DDR model has several limitations that require further improvement. First, although this study showed the recommendation/prediction results by the AC second level due to the sparsity of data, it is necessary to specify the level of recommended drugs to provide more practically useful and helpful recommendations for pharmaceutical companies. In addition, although we have tried to include as many variables as possible regarding the success of drug development based on previous studies to derive meaningful implications on the success of drug development, it will be necessary to add more drug-related variables that can distinguish unique properties by drug. Lastly, we plan to further study on how to combine each algorithm more rationally to build a robust hybrid model. We believe subsequent studies and improvements can readily catalyze the DDR model to advance towards the next step in the near future, such as expansion into product development assistance in the biotechnology.
intelligence era. Nat. Rev. Drug Discov., 19(5), 353–364. https://doi.org/10.1038/s41573-019-0050-3
Schuhmacher, A., Gassmann, O., & Hinder, M. (2016). Changing R&D models in research-based pharmaceutical companies. J. Transl. Med., 14(1), 105.
Schuhmacher, A., Gassmann, O., Hinder, M., & Kuss, M. (2021). The present and future of project management in pharmaceutical R&D. Drug Discov. Today, 26(1), 1–4. https://doi.org/10.1016/j.drudis.2020.07.020
Schuhmacher, A., Gatto, A., Hinder, M., Kuss, M., & Gassmann, O. (2020). The upside of being a digital pharma player. Drug Discov. Today, 25(9), 1569–1574. https://doi.org/10.1016/j.drudis.2020.06.002
Taylor, D. (2016). The Pharmaceutical Industry and the Future of Drug Development. Pharmaceuticals in the Environment (pp. 1–33). The Royal Society of Chemistry.
Thakor, R. T., Anaya, N., Zhang, Y. W., Vilanilam, C., Siah, K. W., Wong, C. H., & Lo, A. W. (2017). Just how good an investment is the biopharmaceutical sector? Nat. Biotechnol., 25(12), 1149–1157. https://doi.org/10.1038/nbt.4023
Tseng, M.-L., Tran, T. P. T., Ha, H. M., Bui, T.-D., & Lim, M. K. (2021). Sustainable industrial and operation engineering trends and challenges Toward Industry 4.0: A data driven analysis. J. Ind. Prod. Eng., 38(8), 581–598.
Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., & Zhao, S. (2019). Applications of machine learning in drug discovery and development. Nat. Rev. Drug Discov., 18(6), 463–477. https://doi.org/10.1038/s41573-019-0024-5
Van Hulse, J., Khoshgoftaar, T. M., & Napolitano, A. (2007). Experimental perspectives on learning from imbalanced data. Paper presented at the Proceedings of the 24th international conference on Machine learning.
Wang, H., Gu, Q., Wei, J., Cao, Z., & Liu, Q. (2015). Mining Drug-Disease Relationships as a Complement to Medical Genetics-Based Drug Repositioning: Where a Recommendation System Meets Genome-Wide Association Studies. Clin. Pharm. Therap., 97(5), 451–454. https://doi.org/10.1002/cpt.82
Zhang, P., Wang, F., & Hu, J. (2014). Towards drug repositioning: A unified computational framework for integrating multiple aspects of drug similarity and disease similarity. AMIA Annu Symp Proc, 2014, 1258–1267.