Journal of Biomedical Informatics 118 (2021) 103792

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Original Research

Explainable artificial intelligence in high-throughput drug repositioning for subgroup stratifications with interventionable potential

Zainab Al-Taie [a,g], Danlu Liu [b], Jonathan B Mitchem [a,c,e,*], Christos Papageorgiou [c], Jussuf T. Kaifi [c,e], Wesley C. Warren [a,c,f], Chi-Ren Shyu [a,b,d,*]

[a] Institute for Data Science & Informatics, University of Missouri, Columbia, MO 65211, USA
[b] Electrical Engineering and Computer Science Department, University of Missouri, Columbia, MO 65211, USA
[c] Department of Surgery, School of Medicine, University of Missouri, Columbia, MO 65212, USA
[d] Department of Medicine, School of Medicine, University of Missouri, Columbia, MO 65212, USA
[e] Harry S. Truman Memorial Veterans' Hospital, Columbia, MO 65201, USA
[f] Department of Animal Sciences, Bond Life Sciences Center, University of Missouri, 1201 Rollins Street, Columbia, MO 65211, USA
[g] Department of Computer Science, College of Science for Women, University of Baghdad, Baghdad, Iraq

ARTICLE INFO

Keywords:
Drug repositioning
Explainable AI
Subgroup stratifications
Data mining
Network analysis

ABSTRACT

Enabling precision medicine requires developing robust patient stratification methods as well as drugs tailored to homogeneous subgroups of patients from a heterogeneous population. Developing *de novo* drugs is expensive and time consuming with an ultimately low FDA approval rate. These limitations make developing new drugs for a small portion of a disease population unfeasible. Therefore, drug repositioning is an essential alternative for developing new drugs for a disease subpopulation. This shows the importance of developing data-driven approaches that find druggable homogeneous subgroups within the disease population and reposition the drugs for these subgroups. In this study, we developed an explainable AI approach for patient stratification and drug repositioning. Contrast pattern mining and network analysis were used to discover homogeneous subgroups within a disease population. For each subgroup, a biomedical network analysis was done to find the drugs that are most relevant to a given subgroup of patients. The set of candidate drugs for each subgroup was ranked using an aggregated drug score assigned to each drug. The proposed method represents a human-in-the-loop framework, where medical experts use the data-driven results to generate hypotheses and obtain insights into potential therapeutic candidates for patients who belong to a subgroup. Colorectal cancer (CRC) was used as a case study. Patients' phenotypic and genotypic data was utilized with a heterogeneous knowledge base because it gives a multi-view perspective for finding new indications for drugs outside of their original use. Our analysis of the top candidate drugs for the subgroups identified by medical experts showed that most of these drugs are cancer-related, and most of them have the potential to be a CRC regimen based on studies in the literature.

1. Introduction

De novo drug discovery is a time-consuming, high-cost, and high-risk process. Developing and implementing a new drug can take anywhere between 10 and 15 years while costing roughly $1.6 billion. The success rate for new drug development is about 2%, with approval rates by the Food and Drug Administration (FDA) declining since 1995 [1,2]. This highlights the necessity of drug repositioning (DR), or the ability to reposition existing FDA approved therapeutics for the treatment of additional diseases [3]. DR takes advantage of existing drug therapies already in use and/or at the approval stage to be declared safe for human administration by the FDA [4]. DR reduces the time, cost, and risk associated with the developmental phases of a new drug application, or (N.D.A.), and represents an important strategy for improving patient care.

Drug-repositioning methodologies involve both computational and experimental techniques. Computational algorithms represent a significant opportunity for the systematic screening and identification of new indications for existing drugs [5–7]. The majority of computational analysis components can be grouped into three categories: machine learning, network analysis, and neuro-linguistics and language semantics [8,9]. Drug repositioning using these methods has been undertaken

* Corresponding authors.
E-mail addresses: mitchemj@health.missouri.edu (J.B. Mitchem), shyuc@missouri.edu (C.-R. Shyu).

https://doi.org/10.1016/j.jbi.2021.103792
Received 30 August 2020; Received in revised form 26 March 2021; Accepted 21 April 2021
Available online 26 April 2021

using disease-centric approaches, drug-centric approaches, or combinations of both [10]. In disease-centric approaches, a drug developed for one disease is suggested for another disease after clustering diseases by phenotypic similarity, molecular signatures, and genetic variation [11–14]. Drug-centric approaches accomplish repositioning based on the similarity of drug molecular activity [15,16]. Some methods are a combination of these approaches based on building drug-drug and disease-disease similarity networks. They then assign drugs based on a *meta*-path score, predicting disease-drug association [17–19], or the correlation between the gene expression profile of a disease and the genes impacted by a drug [20]. Other methods reposition drugs based on mutations, the expression profile of genes, and protein interactions in the diseases of interest [21]. These methods deal with the broad picture of directing a drug to a new disease, but miss the details represented by the response to these drugs on a subpopulation level. The fact that people with the same disease experience different responses to the same drug highlights the importance of looking more deeply into the details of patient subgroup regimens.

Patient stratification into subgroups is a crucial step toward applying precision medicine. As we move to a wider implementation of precision medicine and N-of-1 trials in our health care system [22], it becomes necessary to direct drug discovery in a more patient-centric direction. However, significant barriers exist for the development of new drugs for a small proportion of patients due to the excessive development cost and financial burden to patients. Therefore, DR represents an essential alternative strategy for developing new drugs for patient subpopulations. Once these subgroups are identified, we can more specifically align patients and medications, achieving precision-based therapy. In this work, we developed a novel network-based approach for subgroup drug repositioning. Patients' phenotypic and genotypic data was utilized in conjunction with a heterogeneous knowledge base to provide the most accurate depiction of living systems and their complexity. This heterogeneity gives a multi-view perspective for finding new indications for drugs outside of their original use.

In this study, we apply a novel patient subgroup stratification method to approach drug repositioning by strategically searching a combinatorial phenotypic space with significant genotypic patterns using a biomedical DR knowledge base [23]. This is a unique, network-based, and explainable computational data mining approach for drug repositioning discovery. A heterogeneous knowledge base was adopted from the 'hetionet' project to create our DR knowledge base (DR-KB) [24]. As a case study, we used colorectal cancer (CRC) cases from The Cancer Genome Atlas (TCGA). This dataset contains a large volume of clinical, pathologic, and molecular features allowing us to create highly granular patient subgroup DR recommendations. Our DR approach represents an effective, applicable, and significant opportunity to approach precision medicine using an explainable and data-driven computational method.

2. Related work

The importance of repositioning drugs to tailor treatment for homogeneous subgroups within a heterogeneous disease population has been demonstrated previously by other groups. Most of the existing methods to stratify patients into disease subcategories are based on clinicopathologic features, with some malignancies having shifted towards molecular subtypes [25–28]. The primary method has been to identify drug repositioning candidates for subgroups of patients based on targeting particular genes proven to have a role in disease development. Gouravan, et al. [29] demonstrated that drugs could be repositioned for subgroups of sarcoma patients with well-known mutations that frequently occur, such as finding candidate drugs targeting a BRAF mutation. Simon, et al. [30] focused on a mutation in the RUNX1 gene and studied drug sensitivity to identify candidate drugs for repositioning in patients with acute myeloid leukemia (AML) and a mutation in this gene. Yoshida, et al. [31] showed studies that focus on Myc mutation and investigated this gene family's therapeutic potential across different cancer types. Another method is stratification based on known genotypic variations. After identification, the critical genotypic characteristics of each subgroup are used to identify drugs and targets for repositioning. An example is the repurposing of subtype-specific drugs for breast cancer after the identification of three different modules of Triple-negative breast cancer (TNBC) based on protein-protein interaction networks [27]. Nepal, et al. [32] stratified Intrahepatic cholangiocarcinoma (iCCA) patients based on mutations in three classifier genes, IDH, KRAS, and TP53, and studied their ability to induce substantial downstream molecular heterogeneity and pharmacogenomic potential.

Patient stratification has also been achieved by clustering patients based on a set of gene mutations. Lind and Anderson [33] used machine learning to predict the activity of small-molecule drugs against cancer cells using mutations in oncogenes. In this study, patients were clustered based on their mutation profiles in specific oncogenes and drugs selected targeting these mutations. Gligorijević, et al. [34] also applied machine learning methods to identify patients based on mutation and drug data along with molecular interactions to reposition drugs based on targets in each patient cluster. Additionally, data mining has been used to stratify patients based on molecular features and to prioritize drug targets for repositioning within pre-defined molecular subgroups. Chen and Xu [26] developed a computational method to repurpose drugs for glioblastoma molecular subtypes using human cancer genomics combined with mouse phenotype data.

Though these methods represent a significant step in a promising direction, discovering drugs for a large number of subgroups of patients requires a more comprehensive exploratory approach where multiple factors are considered to address the challenge of patient diversity. The methods outlined above, among others, have shown the importance of patient stratification and have produced promising results. However, patient stratification and DR strategies focused on subgroups with commonly known genotypic characteristics may miss the importance of phenotypic characteristics during the stratification process. Using these methods to solve biomedical problems that impact human life required the implementation of the Explainable AI concept, where the system offers humans the ability to analyze and understand its action and the reasons behind any prediction in order to overcome the black-box challenge of AI in medicine [35]. The Explainable AI concept focuses on transparency, meaning that the actions that affect human life should be explained in a format that is understandable by humans and should show the underlying phenomena of any prediction [36]. The importance of developing explainable computer-based systems to solve biomedical problems has been expressed in studies that include prediction based on medical image processing and genomics analysis [37,38]. Some studies addressed the need for explainable biomedical systems by building interactive ML (iML) models [39,40]. The goal of iML is to enable algorithms to explain each step to users and enable them to correct the provided explanation [41]. Still, the explanations provided by these methods require further study [42]. For patient stratification and drug repositioning, the explainability of machine learning and data mining results can be improved by including the ability to provide insight regarding the underlying biological mechanism that is unique to a subgroup as well as how the perturbation of biological entities contributes to drug selection for a given subgroup. The proposed method aims at building an explainable AI system using data mining and network analysis. This is a step toward building advanced Explainable AI systems that imitate the human cognitive system where the humans make sense of the world by recognizing patterns. Our method provides an exploratory stratification process through the investigation of not only the phenotypic inclusion and exclusion criteria, but also the genotypic characteristics of subgroups that differentiate them from the larger disease population, as well as druggability based on these characteristics. Moreover, our method provides the flexibility to stratify a patient to multiple subgroups. This gives medical practitioners the

ability to consider alternative treatments that remain specific for each patient.

3. Materials and methods

As depicted in Fig. 1, our Patient Stratification and Drug Repositioning (PSDR) framework is composed of three modules. (1) Materials and Data Preprocessing: Patients' genotypic and phenotypic data was

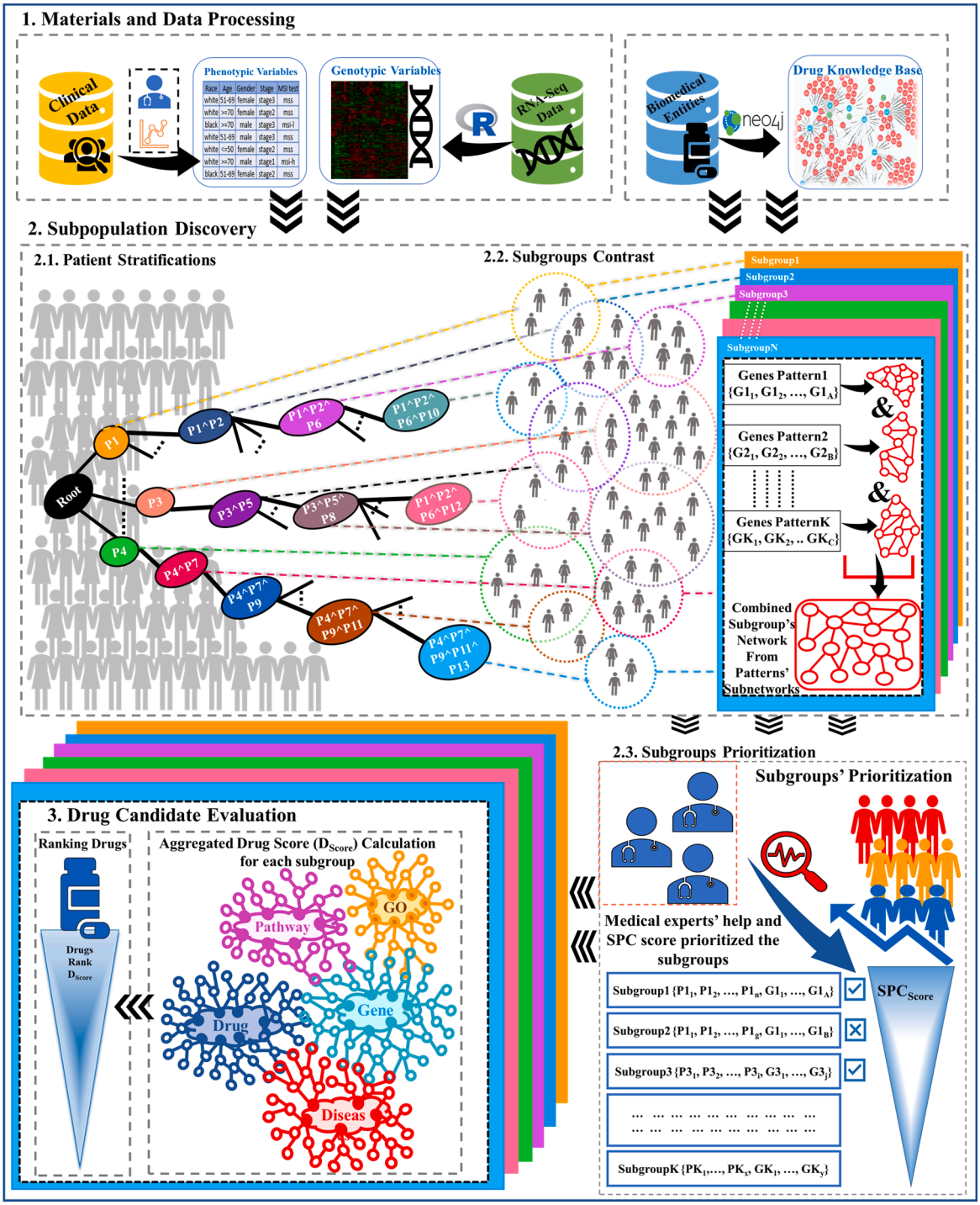

Fig. 1. Patient Stratification and Drug Repositioning Framework. Module 1: The input data consists of patient phenotypic and genotypic characteristics or variables and the DR-KB. Module 2: The subpopulation discovery and evaluation process. Module 2 consists of 3 submodules. Module 2.1: During the path expansion process in which we add or delete a node, module 2.2 is applied to evaluate the contrast and identify the significance of adding or removing a node. Module 2.2: This module is used to calculate the contrast score for each candidate subgroup with the outer population by applying contrast pattern mining and network differentiation. Module 2.3: The subpopulation contrast score (SPC_{score}) is used to rank the candidate subgroups. Medical experts conduct further evaluation of the subgroups that have more potential in clinical settings. Module 3: It is for drug evaluation. In this module, an aggregated drug score (D_{score}) is calculated for each drug in each subgroup network that is created from all the contrast pattern subnetworks. This score is used to rank the drugs for each subgroup to connect the most relevant drugs to a given subgroup.

preprocessed and categorized to be the input to the patient statification algorithm. In this module, a heterogeneous KB was integrated with the patients' data (Section 3.1). (2) Subpopulation Discovery: An explainable AI method for drug repositioning and subgroup discovery of a disease population was developed. For each subgroup, a heterogeneous network was created based on the phenotypic characteristics and gene signatures (Section 3.2). (3) Drug Candidate Evaluation: For each subgroup resulting from the subpopulation discovery module, a drug score was calculated within each network and ranked for prioritization and recommendation of repositioning for each identified subgroup (Section 3.3).

3.1. Materials and data processing

The input data for the patient stratification framework consists of genotypic and phenotypic variables for a disease population. The phenotypic, genotypic, and heterogeneous biomedical networks are used to guide subgroup discovery and recommend drugs for these subgroups. In this study, the genotypic and phenotypic data for CRC patients were obtained from TCGA. The total number of patients included in this study was 565. As part of the human-in-the-loop process, a physician panel involved in the care of CRC patients selected the phenotypic and clinical variables to be included in the analysis. Sixteen clinical variables were selected for relevance to this study. Additionally, most of these phenotypic variables were continuous, which required stratification into categories for inclusion in the data mining algorithm. The physician panel reviewed and approved the categorization of all continuous variables. For example, the original data set has the exact age for each patient. With the help of physicians, we categorize age groups into <50, 50–69, and >=70 for CRC studies. Table S1 in Supplement 1 lists details for the categorization of clinical variables.

The genotypic data in this study are genes differentially expressed between normal and tumor tissues. The differential expression analysis using edgeR was implemented on the RNA-seq data of CRC patients. The dimensionality reduction was made by deciding the p-value to be less than 0.05. $\log2FC > 4$ for the overexpressed and $\log2FC < -4$ for the under-expressed genes. This analysis resulted in 573 significant, differentially expressed genes. Each gene represents a genotypic variable. In addition, a neo4j graph representation of different biomedical entities and the relations between them was used as our DR knowledge base (DR-KB), which contains 11 different types of biomedical variables (Gene, Biological process, Cellular Component, Molecular function, Pathway, Anatomy, Drug (Compound), Side effect, Pharmacological class, Disease, and Symptom) from hetionet [24].

3.2. Subpopulation discovery

In this section, we describe the stratification process based on clinical and genomic characteristics to enable drugs to be directed to a selected list of homogeneous groups within the heterogeneous disease population.

3.2.1. Patient stratification

Finding a homogeneous subgroup cohort is a crucial step in enabling precision medicine. In this study, the focus was to systematically and strategically group CRC patients into phenotypic subgroups based on their genotypic characteristics. We extended our exploratory data mining [23] method by integrating network analysis to guide the subpopulation discovery process. This exploratory data mining method provides an automatic subpopulation discovery tool that computationally investigates a large pool of subpopulations that have underlying factors differentiating each subpopulation within a given larger group. The results are presented as subgroups, each defined by a set of population criteria and underlying factors which differentiate each subgroup from the whole population. These criteria are the phenotypic variables, such as gender, age, and cancer stage. An example subgroup could be males aged less than 50 years with stage II CRC. Each time a phenotypic feature is added, a focus subpopulation is created and contrasted with the rest of the population. Adding additional population variables is desired, assuming there is statistical evidence to do so. The assessment of determining the significance of a subgroup is based on underlying factors which are the genotypic patterns that are statistically unique to the subgroup in comparison to the rest of the population by utilizing Contrast Pattern Mining (CPM) [43].

The patient subgroup stratification module takes a three-level approach. The top-level method, *path expansion*, includes a large number of second-level *floating subgroup selection* processes, each of which is supported by a series of third-level *Inclusion and Exclusion* procedures. This method is exploratory and different from a decision tree approach, where samples are divided based on the decision for each node, and each leaf node contains a group of samples which are exclusively in a particular node. Unlike a traditional decision tree, our approach has a large number of dynamic fanouts for each node without dividing the samples during the expansion process, and each node represents a subgroup. As a result of the patient subgroup stratification process, a patient could be in multiple subgroups through branching expansion. For example, female patients could be grouped into (female, stage III) and (female, age < 50) which potentially contain patients in both groups. To better understand the complexity of this method, let n_p be the number of phenotypic (population) variables with an average n_c categories per variable. There are more than $n_p{}^{n_c}$ subgroups in the exploratory space, which is an unmanageable scale. Therefore, the core of the patient subgroup stratification module (Fig. 1-Module 2) is to automatically and efficiently identify a large number of viable patient subgroups using phenotypic variables, where unique and qualifiable genotypic characteristics are shared by the majority of the patients in the subgroup. This approach differs from traditional greedy algorithms in two perspectives: (1) path expansion selects top potential subgroups that have equivalent performance in druggability of the best and local optimal selection at any stage of the process to ensure the broadness of selected subgroups that are equally viable, and (2) a series of inclusion and exclusion criteria of phenotypic variables using the concept of floating selection approach [44] are performed to avoid simple greedy selection of subgroups.

For each single path in Fig. 1-Module 2.1, the algorithm begins by choosing a single phenotypic variable $(P_i = C_i)$ as a base subgroup, which has the most significant contrast against the rest of the population. The contrast is measured based on the genotypic patterns and subgroup network differentiation from the outer population. Genotypic results guide the algorithm to perform the next inclusion or exclusion of population variables. The subgroup is identified in the stratification process as a group of patients who have the same phenotypic features and share common genotypic patterns and network perturbation patterns, which are unique to that subgroup as compared to the rest of the population. The contrast calculation will be explained in detail in Section 3.2.2. During the floating subgroup selection process, the inclusion step (INCLUSION() function in the pseudo-code and SG2 and SG3 in Fig. 2) adds a new phenotypic variable $P_j = C_j$ to the previous subgroup to generate a more focused subgroup $(P_i = C_i \hat{} P_j = C_j)$. After each inclusion step, the exclusion step (EXCLUSION() function in the pseudo-code and SG5 and SG9 in Fig. 2) is adopted to exclude a less significant move made previously by removing a variable from the "greedy" subgroup if the result of the exclusion process has better performance. For example, when the subgroup is $(P_i = C_i \hat{} P_j = C_j \hat{} P_k = C_k)$ at the third inclusion step, the exclusion step will remove the previous less significant move $(P_i = C_i)$ from the current subgroup if the newly generated subgroup $(P_j = C_j \hat{} P_k = C_k)$ has better "druggability gain" (Fig. 2, 2nd path in the red expansion process). The druggability measurement in this work is computed by the potential drug targets using unique genotypic patterns in the current subgroup contrast with the remainder of the population. The patterns of genotypic variables are extracted using the

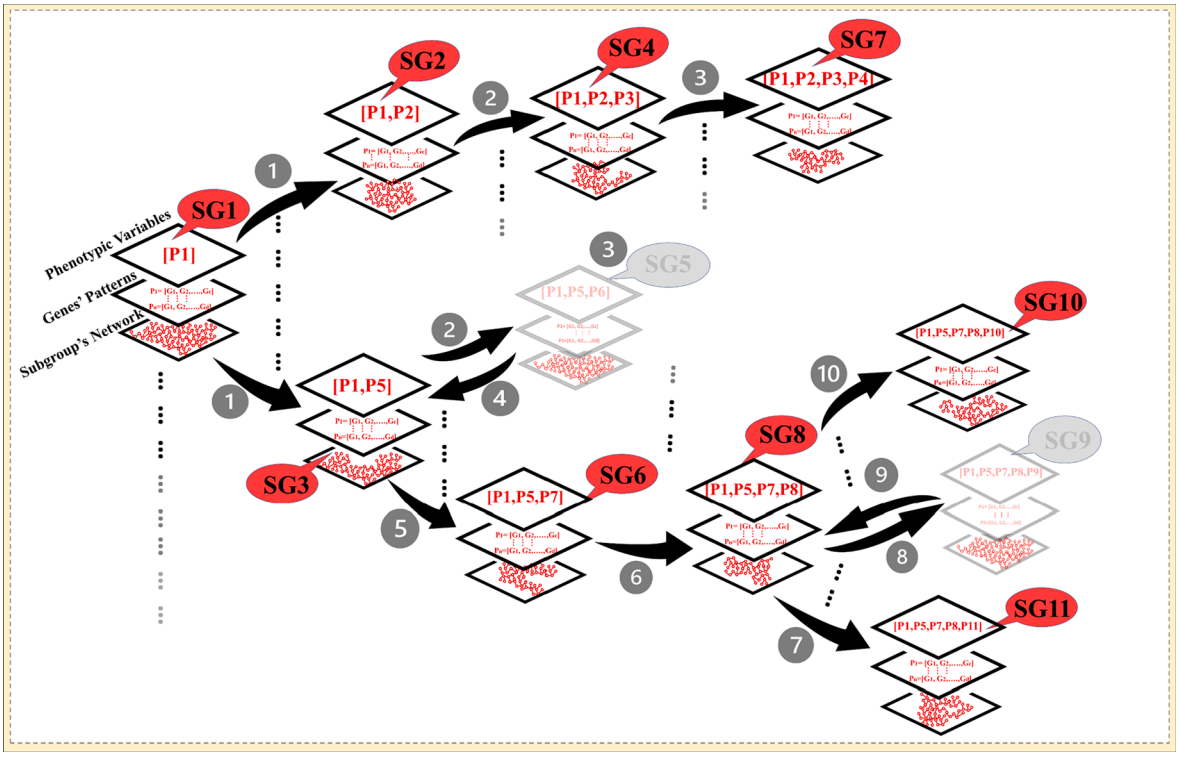

Fig. 2. The subpopulation discovery and evaluation process. It is the floating and path expansion process, where we have different starting points on different computational nodes. For each point, we add or delete a node based on the contrast score. The contrast is evaluated by identifying the significance of adding or removing a node. The contrast score (SPC) is calculated for each candidate subgroup comparing with the outer population by applying contrast pattern mining and network differentiation. Each point represents a potential subgroup, and three layers represent it. The first layer is the phenotypic variables. The second layer is the genes' pattern that are frequent in that group of patients. The third layer is the biomedical interaction of the patterns in the 2nd layer after mapping them to the DR-KB to create the subgroup's network. The SPC score is calculated after comparing the subgroup's network with the rest of the population.

PATTERN_MINING() function, where the algorithm selects the genotypic patterns that frequent in the most patients in a given subgroup. The knowledge base is queried to create a heterogenous network of biomedical entities that interact with these patterns including the protein–protein interactions using the NETWORK_CREATION() function.

Algorithm: Subgroups discovery and drug repositioning
Inputs:
P(D): The phenotypic variable set for dataset D.
SPC(k): Contrast score for subgroups with k variables.
α: Stopping criteria.
M = maximum number of phenotypic variables for contrast subgroups.
Output:
Resulting subgroup with highest contrast SGI
Recommended drugs for SGI
Start
1: $SGI \leftarrow \varnothing$; $k \leftarrow 0$; $SPC(k) \leftarrow 0$;
2: WHILE $k < 2$ DO:
3: Inclusion (P(D), $SCGI$)
4: $k \leftarrow k + 1$
5: END
6: While ((SPC (k) – SPC (k-1))/SPC(k)) > α AND $k < M$ DO:
7: $P_{inclusion}$ = INCLUSION (P(D), SGI)
8: $P_{exclusion}$ = EXCLUSON (P(D), SGI)
9: IF ($P_{inclusion} = P_{exclusion}$) THEN
10: $k \leftarrow k + 1$
11: $SPC(k) \leftarrow SPC(SGI)$
12: ELSE
13: GO TO LINE # 8
14: END
15: DRUG_REPOSITIONING(SGI)
End
Function: INCLUSION (P(D), SGI)
1: SGS: potential subgroup set.
2: $SGS \leftarrow \varnothing$

(*continued on next column*)

(*continued*)

3: FOREACH phenotypic variable $P_i \in P(D)$ DO
4: $CP_{Set}(P_i) \leftarrow [Pairs(\forall categoricalvalue(P_i) \leftrightarrow outerpopulation)]$
5: FOREACH CP($C_{i,a}$, _) $\in$ CP$_{Set}$(P_i)
6: $SGI_{temp} \leftarrow SGI$ + CP($C_{i,a}$, _)
7: $SG_1 \leftarrow D(C_{i,a})$
8: $SG_2 \leftarrow D - D(C_{i,a})$
9: $PTR_1 \leftarrow PATTERN\_MINING(SG_1)$
10: $PTR_2 \leftarrow PATTERN\_MINING(SG_2)$
11: $NW_1 \leftarrow NETWORK\_CREATION(PTR_1)$
12: $NW_2 \leftarrow NETWORK\_CREATION(PTR_2)$
13: SPC(SGI_{temp}) $\leftarrow CONTRAST\_CALCULATION(NW_1, NW_2)$
14: Add SGI_{temp} to SGS
15: END
16: END
17:$SGI_{highest} \leftarrow$ *ThehighestSPCsubgroup*
18: $SGI \leftarrow SGI_{highest}$
19: Remove phenotypic variables of $SGI_{highest}$ from P(D)
Function: EXCLUSION (P(D), SGI)
1: SGS: potential subgroup set.
2: $SGS \leftarrow \varnothing$
3: FOREACH CP($C_{i,a}$, _) $\in CP_{Set}(P_i)$
4: $SGI_{temp} \leftarrow SGI$ – CP($C_{i,a}$, _)
5: $SG_1 \leftarrow D(C_{i,a})$
8: $SG_2 \leftarrow D - D(C_{i,a})$
9: $PTR_1 \leftarrow PATTERN\_MINING(SG_1)$
10: $PTR_2 \leftarrow PATTERN\_MINING(SG_2)$
11: $NW_1 \leftarrow NETWORK\_CREATION(PTR_1)$
12: $NW_2 \leftarrow NETWORK\_CREATION(PTR_2)$
13: SPC(SGI_{temp}) $\leftarrow CONTRAST\_CALCULATION(NW_1, NW_2)$
14: Add SGI_{temp} to SGS
15: END
17:$SGI_{highest} \leftarrow$ *ThehighestSPCsubgroup*
18: $SGI \leftarrow SGI_{highest}$
19: Add phenotypic variables of $SGI_{highest}$ back to P(D)

In this work, instead of tracking only one path, the path expansion process considers the top β% paths based on the druggability gain measurements as the potential successful subgroups at each inclusion and exclusion step, where β is defined as a tracing factor that expands the search to top performers. For example, in parallel to the selection of *P1* at the root of the tree structure in Fig. 2, there are multiple paths that are among the top 10% in druggability performance. This fanout number could range between 1 and $0.1*\sum_{p=1}^{n_p} C_p$ branches per node, where C_p is the number of categorical values of phenotypic variable p. The evaluation of the druggability gain at each step is done by applying subgroups contrast (Section 3.2.2). Unlike traditional optimization methods that attempt to have a sub-optimal solution, this deep mining process will generate a sizable number of subgroups through the expansion process and less greedy results through the floating process. This process is deep mining because the algorithm does not simply offer a single model through a traditional greedy approach. Rather, it takes a sizable number of branches during each step and the final export often generates hundreds to thousands of subgroups. It is analog to deep learning methods to scale the number of encoding and decoding layers with a massive number of neurons that cannot be trained without today's computation power. Our method works in a deep and wide manner to find subgroups. It is "deep" because the algorithm goes from a most general subgroup into a more specific subgroup for each path. This happens by going deeper in each path. The exploratory search will be terminated in each path when the algorithm gets into a most highly specific subgroup with the highest contrast score that cannot be improved further. This algorithm works in a "wide" manner because it explores a large number of "equally creditable" paths from each stratification decision to identify new subgroups.

The subgroup prioritization method (Section 3.2.3) is used to decide whether to keep a node (feature) or remove it if the parent node has more significance than the child node, meaning that the child node does not add further specificity, thus ending that path. We have reported the general algorithmic details of this floating and path expansion process for various biomedical applications [23]. We utilized a distributed computing framework with Apache Spark to run this computationally expensive process. The Big O for the algorithm is $O((\beta n_c n_p^2)^{n_c})$, where β is the tracing factor, n_p is the number of phenotypic (population) variables, and n_c is the average number of categories per variable. After finishing the floating and expansions, the candidate subgroups are prioritized using an index that evaluates the aggregated contributions of all the extracted contrast patterns within each subgroup based on the number of contrast patterns (e.g., co-occurring mutated genes) and the significance (e.g., druggabilities) of those patterns. The final output of our tool is a ranked subgroup list. For each subgroup, we provide the contrast patterns which differentiate a given subgroup from the entire patient population. These patterns provide insight into underlying differences among subgroups and are valuable for further study or clinical trials.

3.2.2. Subgroup contrast

As discussed in the previous section, the evaluation of subgroup significance is performed by measuring the contrast between the subgroup and its outer population. For each candidate subgroup representing a set of phenotypic characteristics, the algorithm finds all genotypic patterns that are frequent within the subgroup but infrequent in the remaining population. Support [45] is used to evaluate whether a given pattern is frequent in a subgroup and growth rate [43] to evaluate the contrast of the pattern in the selected subgroup. In addition, each pattern is evaluated based on its druggability. By mapping these patterns to the DR-KB, the contrast of each subgroup with the outer population is evaluated using multiple biomedical entities, including gene, biological process, cellular component, molecular function, pathway, Anatomy, side effect, pharmacological class, disease, symptom, and drugs that are connected to these patterns in the DR-KB network. Because there are multiple patterns in each subgroup, an overall evaluation of the subgroup can be assessed by aggregating the contributions of the selected patterns. This subgroup evaluation is used to assess the druggability gains on the floating and expansion process in Section 3.2.1.

Let D be the patient dataset in a subgroup, which includes n genotypic variables, $G = (g_1, g_2, \ldots, g_n)$. Pattern p that is commonly shared within patients in a given subgroup is defined as a set of genotypic variables, such as $p = (g_{1,e1}, g_{2,e2}, \ldots, g_{i,ei})$, where $g_{i,ei}$ is the expression level or mutation status of gene i. The expression level or the mutation status should be represented as a categorial value. This process is done using the PATTERN_MINING() function in the pseudo-code.

The pattern is "frequent" if its support is greater than a user-defined threshold. The support of pattern p is the number of records (patients) that have that pattern ($|\langle D,p\rangle|$) divided by the total number of records in the dataset D ($|D|$):

$$Support(p, D) = \frac{|\langle D, p\rangle|}{|D|} \tag{1}$$

To find the contrast pattern (cp) between the focus subpopulation and the rest of the population, S_{G1} represents the focused subgroup and S_{G2} represents the rest of the population, where $S_{G2} = D\text{-}S_{G1}$. The support of the contrast pattern should be significantly different between S_{G1} and S_{G2}. Let s_1 be the support of a contrast pattern in S_{G1} and s_2 the support of the same pattern in S_{G2}. The growth is used to measure the difference between the two groups. The growth of contrast pattern cp between subgroup S_{G1} and the rest of the population S_{G2} is defined as follows:

$$\text{Growth}(cp, S_{G1}, S_{G2}) = \frac{Max\{s_1, s_2\}}{Min\{s_1, s_2\}} \tag{2}$$

The growth ratio is normalized to be between 0 and 1 using an extended version of the tanh function [46].

Let α be the threshold for the support and β the threshold of growth rate. To ensure that a cp is frequent and has significant differences between the two groups, the following condition should be held:

$$(\text{Support}(cp, S_{G1}) \geq \alpha \text{ OR Support } (cp, S_{G2}) \geq \alpha) \text{ AND } (\text{Growth } (cp, S_{G1}, S_{G2}) \geq \beta) \tag{3}$$

This condition identifies two sets of contrast patterns CP_1 and CP_2 for the target subgroup and the outer population, respectively. For each contrast pattern cp_n with multiple genotype variables, the subset of the pattern $cp_i \subseteq cp_n$ will be kept when $\text{Growth}(cp_i, S_{G1}, S_{G2}) - \text{Growth}(cp_n, S_{G1}, S_{G2}) > 0$. These selected contrast patterns are utilized to evaluate each subgroup during the floating and path expansion procedure discussed in Section 3.2.1.

For the purpose of drug repositioning, contrast patterns should embed the druggability of the candidate subgroups. For a gene set in a pattern that is frequent in the focus subgroup but not in the rest of population, the DR-KB is queried to extract the biomedical entities that are connected to each gene in the pattern. Each pattern is represented by a subnetwork of the DR-KB. By integrating all frequent patterns, we obtain an aggregated network for a given subgroup. To measure the

significance of the subgroup based on its relevant patterns and druggability, we calculated a contrast score (SPC_{Score}) using the CONTRAST_CALCULATION() function in the pseudo-code. The calculation of this score is based on values of two major components that were multiplied to get the contrast score (Eq. (5)).

The first component of the product in Eq. (5) measures the contrast of the given subgroup based on the genotypic characteristics of the patients within that subgroup, while the second component measures the contrast of the given subgroup in comparison to the outer population on different levels of biomedical entities that are unique to the subgroup. In the first component, T is a parameter related to the population size t, where $T = \{t, 1/t\}$. $T = t$ when a large population is preferred and $T = 1/t$ when a smaller population is preferred, such as a study of a rare disease. M is the average population size of randomly picked contrast subgroups prior to path expansion. J_{org} and J_{avg} are calculated based on the J-value, that is a quantitative index to evaluate the overall quality of a set of contrast patterns in the subgroup. The J-value for each subgroup will be used to prioritize it among all discovered subgroups. This J-value measurement was inspired by the g-index, which is commonly used to evaluate the productivity of a scholar. If a subgroup (scholar) has a set of patterns (articles), the J-index (g-index) is measured by ranking them in decreasing order based on their growth rate (citations) and then by taking the largest number such that the top J contrast patterns (top g articles) have cumulatively received at least J^2 (g^2) scores. The J-value is defined as follows:

$$J^2 \leq \sum_{i \leq J} Growth(cp, SG_1, SG_2) \tag{4}$$

In the second component of Eq. (5), we consider different biomedical entities in addition to patient genotype patterns that are unique in the subgroup of interest. We are also looking to consider only the biomedical entities that are unique to the subgroup but not to the entire disease population. This is the motivation to include the second component of the product in the equation.

$$SPC_{Score} = \left[\left(T^{*}J_{org} + M^{*}J_{avg}\right)/T + M\right]^{*}\left[1 - \left(\sum_{i=1}^{n}\left(\left(E_{i,1} \cap E_{i,2}\right)/\left(E_{i,1} \cup E_{i,2}\right)\right)/n\right)\right] \tag{5}$$

Patient stratification is completed by considering the patients' specific data and a comprehensive biomedical knowledge base. The knowledge base is heterogeneous to represent the different aspects of the human biological system and the prospective effects of drugs on this system. Each component contributes to the biological and druggable meaningfulness for the patient stratification process. Taking all the biomedical similarities and differences into account in deciding the subgroups is essential to arriving at a more comprehensive assessment for the subgrouping. To address the heterogeneity of biological systems in the context of drug repositioning, we extended the subgroup selection and prioritization method [23] by integrating DR-KB network similarity into the contrast score calculation.

Let $BioE$ be the types of the biomedical entities in the network. $BioE$ = {Gene, Biological process, Cellular component, Molecular function, Pathway, Tissue (Anatomy), Drug (Compound), Side effect, Pharmacological class, Disease, and Symptom}. These different biomedical entities are essential for calculating the subgroups' druggability because the drugs' effect is not only dependent on the genes as isolated entities. These genes are part of different biomedical entities, and perturbations of these genes have different impacts through the relationships of and interactions with various biomedical entities. For example, genes with disease [47–49], pathways [50,51], GO [52–54], tissues [55], side effect [56,57], and the pharmacological classes [58] were used in drug repositioning. In our knowledge base, there are 11 possible biological entity types in the network ($n = |BioE| = 11$). The interactions between genes are based on protein–protein interaction but are represented by the gene names that encode these proteins to reduce the complexity. Gene–interacts–Gene edges represent when the protein products of these genes physically interact [24].

$E_{i,1}$ is biomedical entity type i in the focus group's network, $E_{i,2}$ is biomedical entity type i in outer population network, $E_{i,1} \bigcap E_{i,2}$ represents the number of common entities in entity's type i between the focus group and outer population. $E_{i,1} \bigcup E_{i,2}$ represents the number of all possible entities in entity's type i between the focus group and outer population. By dividing the number of common entities by all possible entities, we obtain the similarity score between the two groups. Subtracting the similarity score from 1 gives the percentage of difference (contrast) between the two groups based on the extracted knowledge from drug repositioning knowledge base. SPC_{Score}, which is a product of the two components, represents the contrast score between the focus group and the outer population (Fig. 1- Module 2.2). This method ensures that each subgroup with common phenotypic characteristics is distinct from the rest of the population. It also ensures homogeneity within each subgroup by including patients who have similar genotypic features. As we ensure homogeneity within each subgroup, our method allows a patient to be in multiple subgroups to provide desirable flexibility in the health care setting. Critically, this provides the ability to find alternative treatment options when the first or second line of treatment fails. Therefore, heterogeneity among subgroups should not be enforced. At the same time, the heterogeneity between the more general subgroup and the more targeted subgroup arises on the genotypic level, where having a smaller subset of the population could enable the algorithm to discover new genotypic patterns that were not statistically significant in a more general population. To ensure statistical significance, we kept only the subgroups with a p-value < 0.05. In this study, there were 130 statistically significant subgroups with at least 50 patients in each subgroup.

3.2.3. Subgroups prioritization

The number of candidate subgroups selected by the floating and expansion process could be hundreds. The SPC_{Score} is used to rank the subgroups. The higher the SPC_{Score}, the higher the potential for drug repositioning. Because this method was developed to improve patient care, we ensured that all the steps were explainable and acceptable for the practitioners. For clinically meaningful results, a physician-in-the-loop process was necessary to prioritize the subgroups further using a two-phase method. First, physician-in-the-loop provides a filtering mechanism where the focus will be on only a subset of the subgroups instead of going through the hundreds of subgroups resulting from our method. Second, the physicians may decide the most relevant subgroups by evaluating the top subgroups using the SPC_{Score} or using initial hypotheses formed by clinical observations and literature to prioritize all candidate subgroups. For example, in this study, the physician investigators decided to focus on the subgroups with MS status as one of the clinical variables. This resulted in 25 subgroups. Then, we further examined seven subgroups with Microsatellite Instability (MSI) test results as a phenotypic characteristic among 130 statistically significant subgroups (p-value < 0.05). The rationale for the selection of groups

based on MS status is related to therapeutic selection and tumor biology [59]. Microsatellite instable tumors are associated with hypermutation due to inactivation of mismatch repair genes via either germline mutation or methylation, accounting for 13–15% of CRCs. The remaining 85% of colorectal cancers develop via the chromosomal instability pathway, referred to as microsatellite stable, following a well-described pathway acquiring mutations through the adenoma to carcinoma sequence as described in seminal work by Vogelstein, et al [60]. While these tumors appear to be biologically different, most critically, these tumors are also characterized by different prognosis, response to standard therapy, and response to novel therapy including both targeted and immune-based therapy [61–63]. Therefore, this designation was felt to be highly clinically relevant, and the subgroups that were selected through the physician-in-the-loop process were then chosen as the input for the next step where drug candidates were evaluated and analyzed for each of these subgroups.

3.3. Drug candidate evaluation

The biomedical entities directly and indirectly connected to each gene, which is DEG, as discussed in Section 3.1, in the patterns are extracted from the DR-KB (Fig. 1- Module 2.3), which is represented as a neo4j directed graph $G=(V_G, E_G)$. For drug repositioning, drugs within each subgroup's network need to be evaluated based on each drug's connectivities in the network. This evaluation depends on how many genes in the subgroup are affected by that drug and the connectivity between the genes and other biomedical entities in the network. To accomplish this, an aggregated drug score is calculated over each subgroup's network for each drug.

Let $GS=(g_1, g_2, \ldots, g_n)$ be the genotypic signature of subgroup S_{Gi}. A graph H composed of a collection of sub-networks for all genes in GS can be obtained through the following equation:

$$H = \{(V_H, E_H) | (V_H \subseteq V_G) \wedge (E_H \subseteq E_V) \forall g \in GS\} \quad (6)$$

where V_H is a set of vertices in which g is reachable. The resulting network is representative of the subgroup with different biomedical entities, including genes, biological processes, cellular components, molecular functions, pathways, tissues, diseases, and drugs. Due to multiple candidate drugs in each subgroup's network, the drug prioritization method is needed to prioritize drugs within each network. An aggregated weight calculation algorithm is used to prioritize the drugs in each subgroup's network.

Let G' be the subgroup's network. $G'=(V', E')$ is a rooted graph where all vertices are directed toward the drugs d as shown in Fig. 3. G' has edge weights w: $E' \rightarrow \mathbb{R}$. Assignments of weights are based on interaction types. For example, gene-gene interaction has a higher weight than that of tissue-disease interactions.

In the subgroup network, we kept only the entities that are directly or indirectly connected to genes that, in turn, are connected to drugs. We determined that a gene is connected to a drug if:

- There is a direct edge $e_{ij}(v_j,v_i) \in E'$ that connects gene v_j to drug v_i (gene g_1 and drug d_i in Fig. 3-A).
- There is an indirect edge, $e_{ij}(v_j,v_i)$, that connects one or more genes of the subgroup's genes set to a drug through another gene if $e_{ix}(v_x, v_i) \in E'$ and $e_{xj}(v_j,v_x) \in E'$, where v_x is a gene (g_a in Fig. 3-B).
- There is an indirect edge, $e_{ij}(v_j,v_i)$, that connects a gene to a drug through a disease if $e_{ix}(v_x,v_i) \in E'$ and $e_{xj}(v_j,v_x) \in E'$, where v_x is a disease (s_n in Fig. 3-C).

First, a score is calculated for each gene in G'. The score is the weighted sum of all paths that connect leaf nodes to the given gene. Let P be a path that goes from a leaf node v_k to the given gene v_i, $P=<v_i, v_1, \ldots, v_k>$, and N_{ik} is the total number of paths directed toward node v_i from any leaf node v_k.

$$w(vi) = \sum_{p=1}^{N_{ik}} \sum_{j=1}^{k} w(v_{p,j}, v_{p,j-1}), \quad (7)$$

where $w(v_i)$ is the weight or the score of the given gene v_i and $w(v_{p,j}, v_{p,j-1})$ is the weight of the edge that pointed from node v_{j-1} to the node v_j in path P. A drug score (D_{Score}) can be calculated by calculating the sum of the weights of all genes that have interactions with that drug.

$$DScore = \sum_{n=1}^{d_{G'}^{-}(d_i)} w(v_n), \quad (8)$$

where $d_{G'}^{-}(d_i)$ is the in-degree of vertex d_i in G'. $w(v_n)$ is the weight of a gene with index n within the set of genes that are connected to drug d_i.

In the output, the subgroups discovery and drug repositioning framework returns a list of subgroups. Each subgroup has a contrast score, SPC_{Score}, representing the contrasts between a given subgroup and the entire disease population. For each subgroup, a set of drugs within each subgroup's network are ranked using D_{Score}, where the higher the

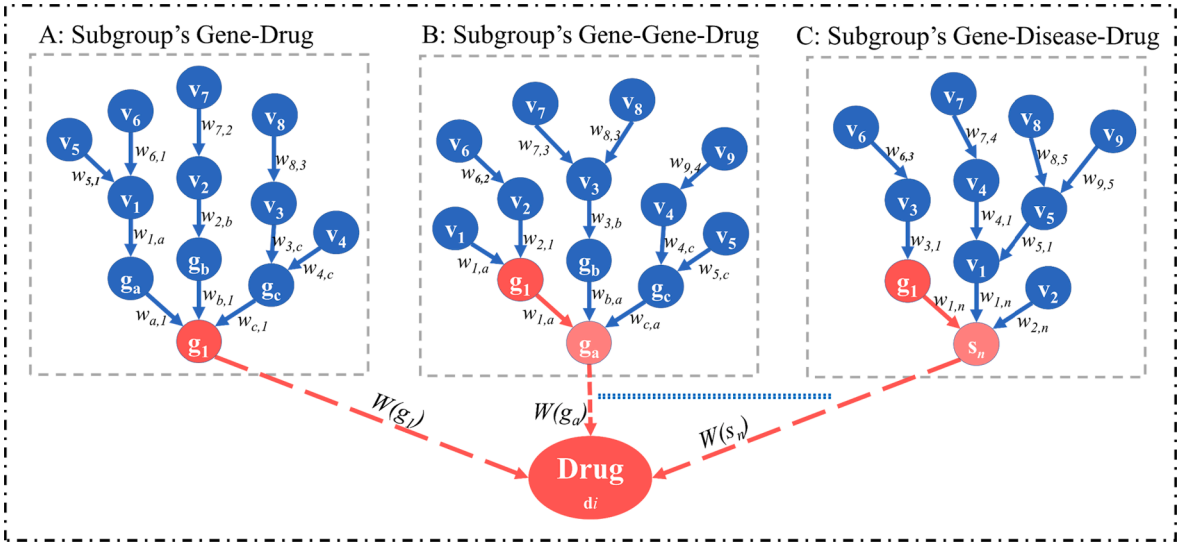

Fig. 3. An example of aggregated drug score calculation for each drug, where v's represent the vertices of that subgroup's network, and w's are the weights of the edges that connect one node to another. The aggregated score is calculated layer by layer from the leaf nodes to the root nodes (g_1, g_a, and s_n). The final drug score is calculated by summing up the aggregated score (w(g's)) of all the genes connected to that drug.

score, the more relevant that drug is to the subgroup. In each stage of the analysis, medical experts, the physician co-authors specializing in breast, colorectal, and lung cancers, were included in the decision process for the next step of the analysis. In the evaluation for the candidate drugs, a physician-in-the-loop is required to evaluate both the effectiveness and side effects of top-ranked drugs. Physicians can assess the effectiveness of a drug based on the relationship between the drug's molecular profile from one side, and the gene patterns in the perturbed biological entities from the other side. Such explainable results, contributed from the motivation and design of the algorithm, are intuitive to clinicians when explaining why the drugs are recommended with underlying biological mechanisms to health care providers, as explainability is a critical limitation to the adoption of many current data mining methods. For the candidate drugs, the physician can further evaluate patient comorbidity, risk factors, and medical history to assess for interactions and potential side effects.

4. Results

Differentially expressed genes and the phenotypic data were used with the heterogeneous biomedical knowledge base for subgroup discovery and drug repositioning using CRC patients from the TCGA. In this section, the results of our analysis are explained using CRC as a case study. Section 4.1 explains the subpopulation discovery results. Section 4.2 concerns drug repositioning and prioritization for each subgroup within this CRC population. Section 4.3 discusses the relations between drugs and genes in the subgroup networks. Section 4.4 presents the analysis of the subgroups' top-ranked drug candidates.

4.1. Subpopulation results

After performing exploratory data mining, we filtered the resulting subgroups based on the SPC_{Score}. Subgroups with $SPC_{Score} > 0$ were considered to be signficant CRC subgroups. The total number of subgroups that met this condition was 130 (see Supplement 2). The SPC_{Score} ranged between 2.5 and 53 (mean $SPC_{Score} = 17.44 \pm 8.60$; see Table 1 and Supplement 2). These subgroups were then categorized based on clinically relevant features in consultation with clinicians experienced in treating CRC patients. Among 130 subgroups, we then chose to focus on a set of subgroups with population variables containing microsatellite status (MS) which has three possible values/categories, namely, microsatellite instability-high (MSI-H), microsatellite instability-low (MSI-L), and microsatellite stability (MSS). Microsatellite status is a critical clinical feature of CRC, as studies have demonstrated important molecular differences impacting treatment response [64]. Additionally, patients with MSI-H tumors are the only patients with CRC that have demonstrated significant responses to immune checkpoint blockade [63]. Using this sub-categorization, we found 25 subgroups with MSI status as part of the population variables listed in Table 1.

In addition to MS status, there are critical differences in treatment outcomes based on specific clinicopathologic factors such as gender, anatomic location, and lymphatic invasion [65–68]. These factors were also found to be critical features using our exploratory data mining algorithm when combined with MS status. Due to clinical significance, we chose to focus on specific subgroups with these critical and clinically relevant features for in-depth study (see Table 1, Table S2 in Supplement 1, and Supplement 3). To pictorially present subgroups from the three categories, Fig. 4 shows subgroups matched with relevant drugs. For example, within the subgroups that have MSI-H as a phenotypic variable, right-sided colon cancer (P2R) and no lymphatic invasion (P6N) are features in three different subgroups. The first subgroup has MSI-H and P2R as phenotypic features and the suggested drugs for this subgroup are Cerulenin, Crizotinib, and Afatinib. The second subgroup has MSI-H and P6N as phenotypic features and the suggested drugs for this subgroup are Idarubicin, Dactinomycin, and Doxorubicin. The third subgroup has MSI-H, P2R and P6N as phenotypic features and the suggested drugs for this subgroup are Menadione, Dasatinib, and Vinblastine.

4.2. Drug repositioning results

Unique differentially expressed genes for each subgroup were used to query the drug knowledge base (see Table 1). For each differentially expressed gene (DEG), all the biomedical entities were retrieved to create a network that represents the given subgroup (see Table S3 in Supplement 1, and Supplement 3).

(1) The patients in this subgroup have **MSI-H** as their MSI test result and have **no lymphatic invasion**. This subgroup contains about 10% of patients. The uniquely differentially expressed genes are RAB37, GCK, and NOP56. The top three predicted candidate drugs for this subgroup are Idarubicin, Dactinomycin, and Doxorubicin.
(2) The patients in this subgroup have **MSI-H** as their MSI test result, and they have **right-sided colon** as their anatomic neoplasm subdivision. This subgroup contains about 14% of the sample. The uniquely differentially expressed genes are PLXNA1, FDXR, AGRN, and MYO7A. The top three predicted candidate drugs for this subgroup are Cerulenin, Crizotinib, and Afatinib.
(3) The patients in this subgroup have **MSI-H** as their MSI test result. These patients have **right-sided colon** as their anatomic neoplasm subdivision with **no lymphatic invasion**. This subgroup contains about 10% of the sample. The uniquely differentially expressed genes are TCEA2, SNRPN, SPG20, LOC157381, YPEL4, MUM1L1, FBXO17, MYLK3, NHLRC1, ELMOD1, COL25A1, CBLN4, LOC339535, SOX30, and KCNJ3. The top three predicted candidate drugs for this subgroup are Menadione, Dasatinib, and Vinblastine.
(4) The patients in this subgroup are **female** with **MSS** as their MSI test result. This subgroup contains about 29% of the sample. The uniquely differentially expressed genes are PRKY, GYG2, XIST, and COX7B2. The top three predicted candidate drugs for this subgroup are Varenicline, Digitoxin, and Gefitinib.
(5) The patients in this subgroup are **female** with **MSS** as their MSI test result, and they have **lymphatic invasion**. This subgroup contains about 12% of the sample. The uniquely differentially expressed genes are CPT2, FNIP2, PANK3, SIPA1L2, PPARGC1B, GCET2, RAB27B, ALDH1A1, EPHA4, SLITRK6, UGT2B7, GAS2L2, KLK3, DEFA5, C1orf112, RPL23AP7, MFRP, NOS3, ARSE, TBX6, TNFRSF4, FCRLB, SUSD3, MYL4, AQP7P1, SNHG9, MMP17, MPO, C10orf82, ART5, NKAIN4, PCDHA4, UPB1, PRINS, PLSCR2, MLANA, PKHD1, C14orf53, ZPBP2, HBG1,

Table 1
Categories of MSI test subgroup s. MSI-H = microsatellite instability-high; MSI-L = microsatellite instability-low; MSS = microsatellite stable.

Subgroup category	Number of subgroups	Number of population variables		SPC_{Score}		Number of patients per subgroup		Genes per group		Number of drugs per group	
		Min	Max	Mean	SD	Mean	SD	Mean	SD	Mean	SD
MSI-H	11	1	3	37.85	10.39	64	12	642.55	518.88	845.27	267.39
MSI-L	2	2	3	4.21	2.47	54	4	1045.5	183.14	1135.5	36.06
MSS	12	1	4	15.22	14.96	130	100	1234	701.28	1125.40	213.80

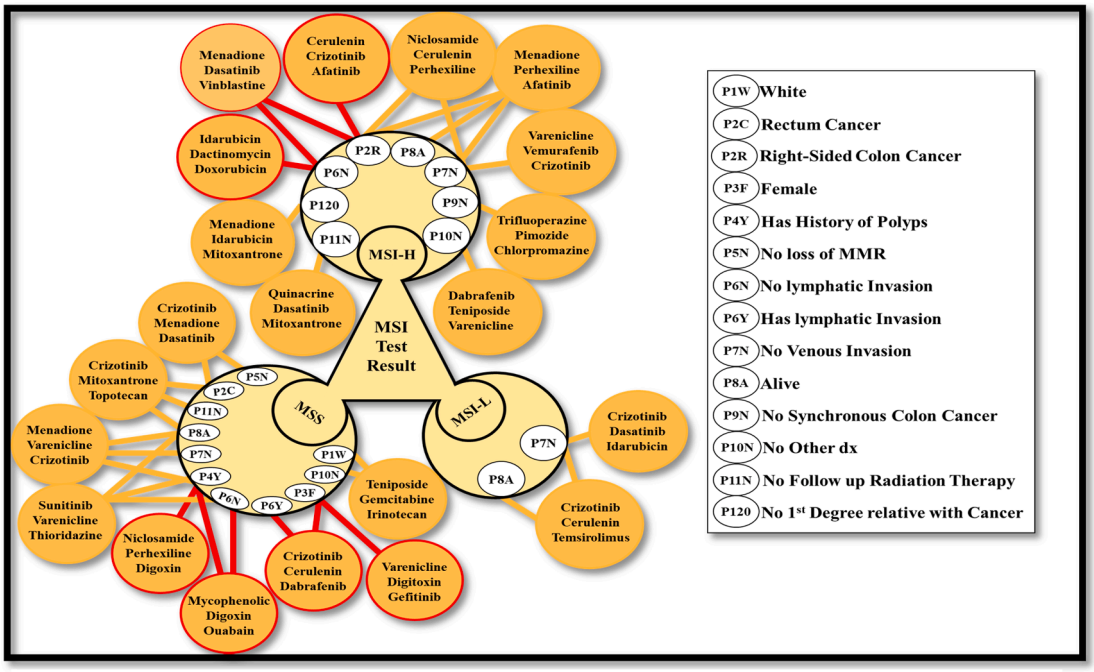

Fig. 4. MSI test result-related subgroups with the top three recommended drugs for each subgroup. The white circles are the phenotypic properties, and the golden circles are the recommended drugs for a subgroup that have the phenotypic properties that are connected to the given drugs' circle in addition to the MSI test result. The phenotypic variables connected to drugs circled with a red border are the subgroups that are highlighted. MSI-H = microsatellite instability-high; MSI-L = microsatellite instability-low; MSS = microsatellite stable; MMR = mismatch repair; dx = diagnosis. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

PCDHA12, KIAA0087, LOC100133469, and PCDHA11. The top three predicted candidate drugs for this subgroup are Crizotinib, Cerulenin, and Dabrafenib.

(6) The patients in this subgroup have **MSS** as their MSI test result, and they have **a history of colon polyps**. This subgroup contains about 20% of the sample. The uniquely differentially expressed genes are CACNG8, OR10H1, LOC440173, DUXA, KRTAP10-6, KCNA10, FGF21, SSX7, CACNG1, KRTAP5-7, HIST1H2BF, GPR144, GOLGA9P, INGX, DSPP, P2RX3, EPX, RNF222, KRTAP10-12, LOC100128675, LCNL1, FAM75A3, KRTAP10-2, NPBWR1, GPR152, FAM75A6, C14orf166B, TAS1R2, SLC22A8, RGS7, PTX4, FLJ42393, and RBMXL3. The top three predicted candidate drugs for this subgroup are Niclosamide, Perhexiline, and Digoxin.

(7) The patients in this subgroup have **MSS** as their MSI test result, have **a history of colon polyps**, and they do **not have venous invasion**. This subgroup contains about 9% of the sample. The uniquely differentially expressed genes are NDRG1, NLGN3, TGFBR3, GABRB3, RPS5, ROMO1, SNAR-C3, and C9. The top three predicted candidate drugs for this subgroup are Menadione, Varenicline, and Crizotinib.

In our analysis, we focus on the top 3 drugs for each of the seven subgroups. The total number of drugs is 16 drugs. Twelve of these 16 drugs are cancer-related therapies. Eight of these twelve are also associated with CRC treatment. These drugs are Dasatinib, Crizotinib, Niclosamide, Dabrafenib, Gefitinib, Afatinib, Doxorubicin, and Menadione. These drugs were used or suggested as either mono or combined therapy for CRC patients with different genetic features.

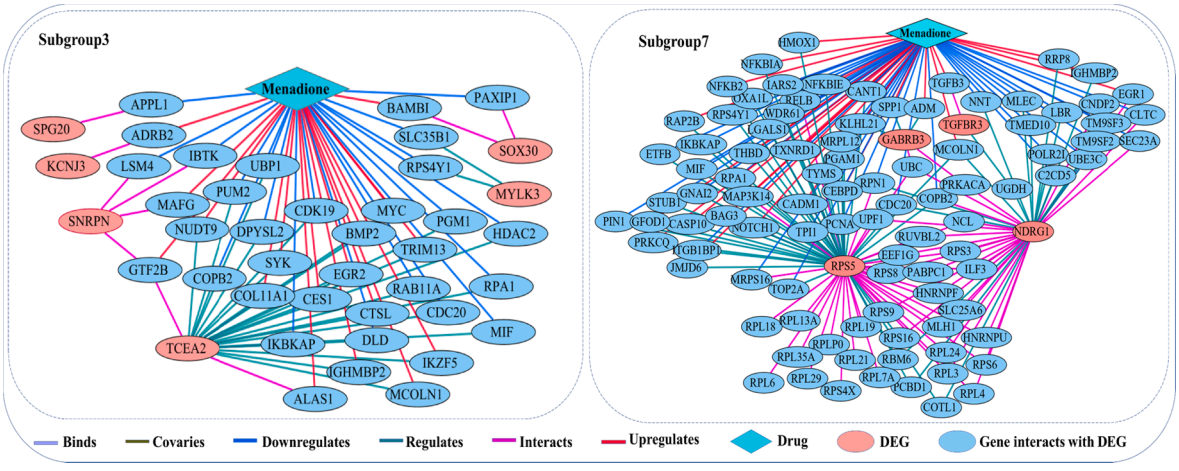

Fig. 5. Menadione gene interactions in two different subgroups.

4.3. Drug-differentially expressed gene subgroup networks

Different biomedical entities in a subgroup network have different roles in deciding the drug and drug ranking within the network. The genes and gene interactions help determine the extracted drugs, while other biomedical entities factor in weighting the genes and ranking the drugs. The analysis for the genes that caused drugs to be suggested for a subgroup showed the importance of pathway and gene interactions included in the study. DEGs in a subgroup interact with different drug targets and genes that are affected by the treatment to create the molecular profile of a drug in the given subgroup. For example, Menadione was suggested in subgroup3 and subgroup7 because it affects genes that have direct interactions with the DEGs in these subgroups. Menadione is a form of vitamin K that has a critical role in blood clotting and bone health. Based on *in vitro* cell line investigations, Menadione was found to have anti-cancer effects, including in CRC [69,70]. In subgroup3, Menadione downregulates 19 genes and upregulates 17 of the genes that interact with six DEGs unique to subgroup3 (Fig. 5). In subgroup7, Menadione downregulates 36 genes and upregulates 25 genes that have direct interaction with six of the DEGs for that subgroup (Fig. 5).

For the top three drugs in these seven subgroups, Crizotinib was the one most frequently suggested (3/7). Crizotinib was recommended for subgroup2, subgroup5, and subgroup7 (Fig. 6). As expressed in DrugBank, Crizotinib is "an inhibitor of receptor tyrosine kinase for the treatment of non-small cell lung cancer (NSCLC)." Some studies have also investigated its potential in CRC patients. When used for combination therapy in MSI-H, BRCA2 deficient patients with c-MET overexpression, Crizotinib was found to increase apoptosis and tumor cell death [71]. In subgoup2, where Crizotinib was recommended for patients who have MSI-H right-sided CRC, MET gene expression was upregulated, and mutated in 28% of patients. For CRC tumors where SOX13 mediates cells migration, invasion, and metastasis, it has been found that inhibiting c-MET using crizotinib prevents CRC metastasis by blocking HGF/STAT3/SOX13/c-MET axis [72] and SOX13 is mutated in 19% of patients in this subgroup. Crizotinib was found to have a role in overcoming the resistance to some drugs like cetuximab. In some CRC cell lines, the resistance was developed to cetuximab as a result of activating Tyrosine Kinases (RTKs) like MET and RON, or the resistance that was a result of adding HGF and NRG. Adding crizotinib to these cell lines blocked resistance to cetuximab [73]. Also, crizotinib has also been shown to overcome resistance to cetuximab and improve the chemoradiation outcome in CRC cell lines that carry mutant KRAS [74]. In subgroup5, where the patients are females that are MSS and have lymphatic invasion, MET is mutated in 15% of patients and gene expression is upregulated. BRCA2 was also upregulated and mutated in 9% of these patients. In subgroup7 where the patients have MSS and a history of colon polyps with no lymphatic invasion, MET was upregulated and mutated in 23% of patients with STAT3 mutated in 9% of patients. All these data support the recommendation of crizotinib for these patients and demonstrate the power of this explainable data mining application.

As we observed with Menadione and Crizotinib, a drug can be recommended for more than one subgroup though the DEGs are different, and the genes set by drug interactions are different in different subgroups. This shows that our method for hypothesis generation can be used to recommend drugs based on the molecular profile for patients within each subgroup when matching the gene signature of the drug and the gene signatures of patients in identified subgroups. Also, when comparing drugs within the same subgroup of patients taking Menadione and Crizotinib as an example, we can see in subgroup7 that there is a set of genes that are the same in both drugs' networks. However, to decide which drugs should be used for which patients within that subgroup, we can find the unique gene signature in each drug's network. In subgroup7, about 78% of the genes in the Crizotinib network are unique to Crizotinib, while in Menadione's network, about 84% of the genes are unique to Menadione's network.

The validation of each drug recommended for each subgroup is explained in Table S3 of Supplement 1, where each row in the table represents the matching between the drug profile and the genotypic

Table 2
The results of the randomized analyses. It lists the top three drugs of our DR method and the top three drugs of the randomized networks for each of the selected subgroups.

	Subgroups	DR Top Drugs	Randomized Top Drugs
1	MSI-H & no lymphatic invasion	Idarubicin Dactinomycin Doxorubicin	Tetrahydrobiopterin Diethylstilbestrol Fluphenazine
2	MSI-H & right-sided colon	Cerulenin Crizotinib Afatinib	Streptozocin Tetrahydrobiopterin Metaxalone
3	MSI-H & right-sided colon & no lymphatic invasion	Menadione Dasatinib Vinblastine	Testosterone Pseudoephedrine Paliperidone
4	MSS & female	Varenicline Digitoxin Gefitinib	Terazosin Isoflurane Fospropofol
5	MSS & female & have lymphatic invasion	Crizotinib Cerulenin Dabrafenib	Imatinib Simvastatin Bortezomib
6	MSS & a history of colon polyp	Niclosamide Perhexiline Digoxin	Trilostane Doxylamine Flucloxacillin
7	MSS & a history of colon polyp & No venous invasion	Menadione Varenicline Crizotinib	Diclofenac Digoxin Progesterone

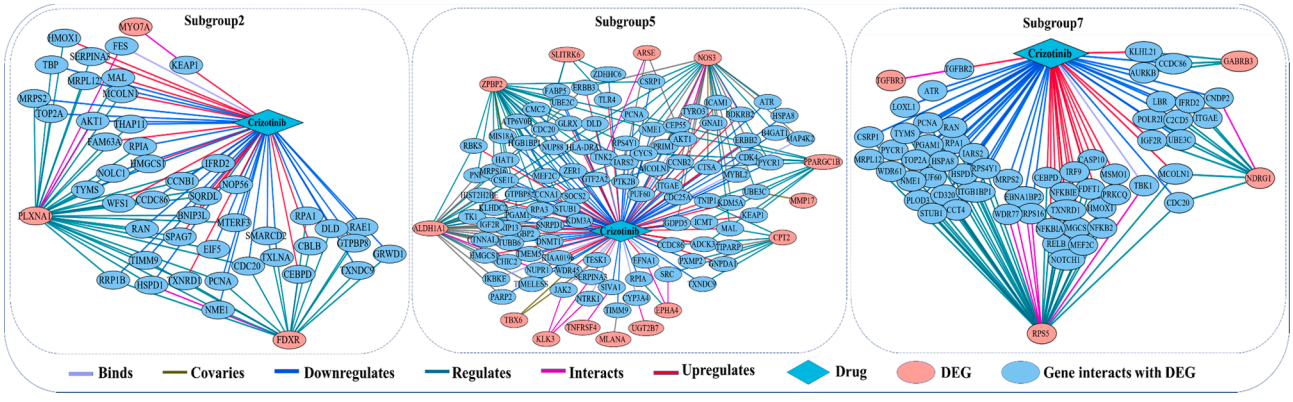

Fig. 6. Crizotinib gene interactions in three different subgroups.

features of the patients in a given subgroup from the literature.

4.4. Analyzing the subgroups' top-ranked drug candidates

4.4.1. Randomized analysis

For validation, a randomized analysis was conducted within each subgroup. For each subgroup's network, we preserved the nodes and shuffled the edges to create a random network. For each subgroup, we created ten random networks. Then, we ran our drug repositioning method on each network. The results showed that the top drugs generated based on the randomized networks were different from the top 15 drugs generated from our original subgroup networks. Indeed, there were only two candidate drugs that were ranked 18th and 34th from top drug list in the randomized network. The remaining drugs were ranked greater than 40. Table 2 shows the top three drugs from both methods in detail. In the table, we show the top 3 drugs for the random networks based on the average rank for each drug in a given subgroup.

4.4.2. Drugs' classes enrichment analysis

The class of each drug among the top 10 repositioned candidates was examined to evaluate the candidate drugs recommended for the seven subgroups. After obtaining the top 10 drugs for each subgroup and removing duplicates, we found 37 unique drugs. Then, we mapped these drugs to the knowledgebase to link the pharmacological class for each. We found 31 different classes were enriched in these drugs. The top five pharamcological classes are shown in Fig. 7. The remaining classes were enriched in less than 5% of the drugs. The result shows that Topoisomerase Inhibitors, Protein Kinase Inhibitors, Anthracyclines, P-Glycoprotein Inhibitors, and Corticosteroid Hormone Receptor Agonists were the pharamcological classes that were highly enriched with the top 10 drugs for the seven subgroups of interest. Topoisomerase Inhibitors, such as Irinotecan, a drug commonly used in the treatment of CRC, are considered some of the most effective apoptosis inducers [75]. This is due to their ability to target Topoisomerase enzymes, which have a significant role in DNA replication [76]. The sensitivity to these inhibitors was also found to be increased in colorectal cancer cell lines defective in DNA MMR, a critical patient group [77]. The second class is protein kinase inhibitors, which have been developed to block pathways related to tumor growth and progression. Studies have shown that these inhibitors have the potential to be used in the treatment of metastatic colorectal cancer [78]. Anthracyclines have been shown to be effective for the treatment of breast cancer with TOP2A mutations, and patients with metastatic CRC were found to have a higher rate of mutation in this gene than in breast cancer, leading to a phase II trial in CRC [79,80]. Different P-Glycoprotein Inhibitors have been developed or recommended as repurposed drugs to treat cancer [81]. This shows that the majority of our drugs belong to cancer-related classes and have the potential to be repositioned for colorectal cancer.

4.4.3. Pathways enrichment analysis

To understand the common underlying mechanism of action for the top drugs from all the subgroups of interest, we analyzed signaling pathways that are highly enriched with these drugs' gene targets. For the top 10 drugs of each subgroup, a gene set enrichment analysis was performed to find the highly enriched pathways for each drug's targets [82,83]. For the 37 drugs, we examined the top 10 pathways for each drug. A summary of our findings is shown in Fig. 8, where the x-axis represents the top pathways that are highly enriched in the repositioned drug candidates, and the y-axis represents the percentage of drugs in which a given pathway was within the top 10 pathways targeted by a drug.

The pathway that was targeted by about 80% of these drugs was the cell cycle pathway; then, the p53 pathways followed by chronic myeloid leukemia. Different drugs were developed and have been shown to produce anticancer effects affecting signaling pathways, including the cell cycle. Drugs like proteasome inhibitors have shown effectiveness on human CRC cells [84]. Another study revealed that affecting cell cycle components inhibits colorectal cancer cell proliferation [85]. The second pathway is the p53 pathway. P53 is a key tumor suppressor gene that is mutated or lost in different cancer types including CRC. Regulating the p53 pathway impacts apoptosis [86]. Also, some studies have suggested that restoring the p53 pathway may enable selective cell death in cancer, including CRC, where the cancer cells can be targetted without affecting the normal cells [87]. The third pathway was chronic myeloid leukemia. A study found that a c-kit tyrosine kinase inhibitor with a significant effect on chronic myeloid leukemia could be repurposed for colorectal cancer patients expressing the c-kit proto-oncogene [88]. In this study, and after testing the inhibitor in human colorectal tumor cells in vivo and *in vitro*, the inhibitor has the potential to prevent colon cancer and treat advanced CRC related to liver metastases. Additionally, the mismatch repair pathway, a common dysregulated pathway in CRC, and the colorectal cancer pathway were highly targeted pathways by these drugs. These results show that the recommended drugs for the subgroups of interest are highly enriched for targeting cancer-related pathways.

5. Discussion

In this study, we presented a framework for patient subgroup stratification and drug repositioning. Its explainable results can be used to

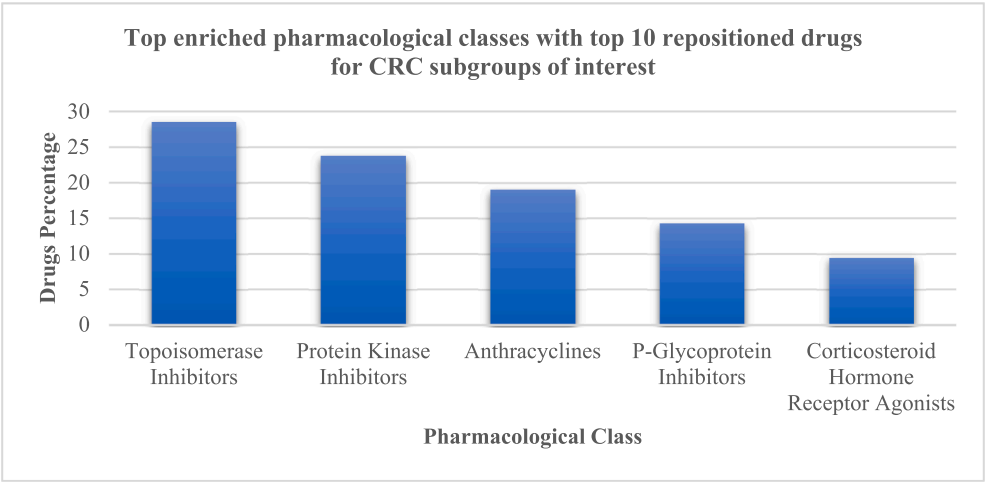

Fig. 7. Top pharmacological classes that were highly enriched with the drugs repositioning candidates for the seven subgroups of interest.

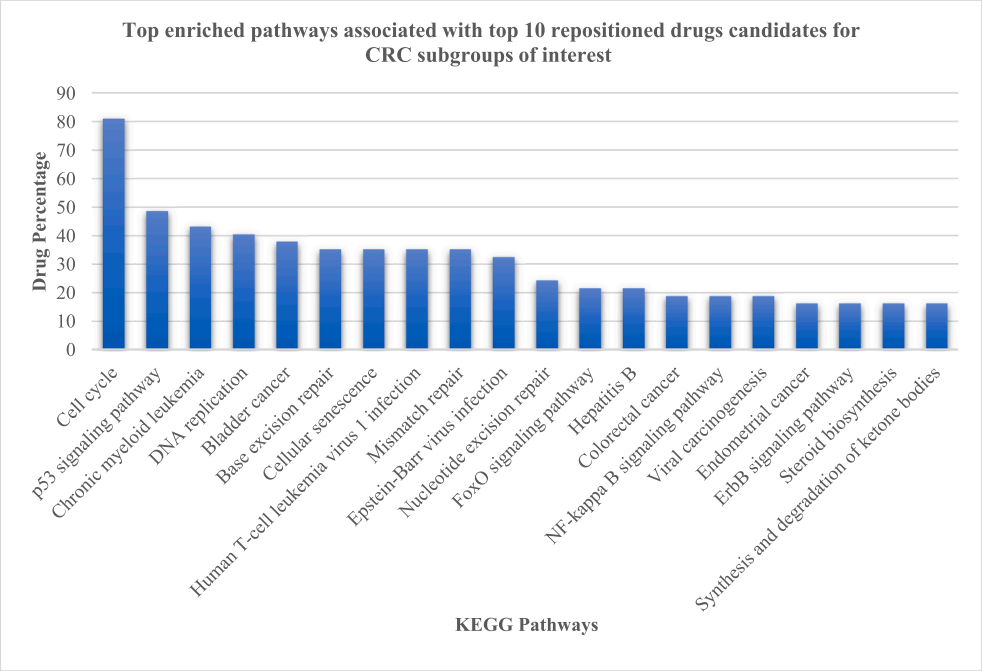

Fig. 8. Top pathways that were targeted by repositioned drug candidates for the seven subgroups of interest.

generate hypotheses for future clinical trials in which researchers and physicians can tailor treatments to subgroups of patients through the automatic determination of inclusion and exclusion criteria and the unique molecular profile of each candidate drug within a subgroup. The subgroups were selected based on phenotypic features and genotypic patterns. The differentiation between a subgroup and its outer population is based on the contrast using these patterns and network analysis. Part of the explainability of our findings is that the SPC_{Score} and D_{Score} can be traced back to the basic biomedical components supporting the selection of the repurposed drugs. The explainability of this method demonstrates the ability to explain the reasoning behind the selection of the subgroups and drugs. As different research has shown the possibility of non-cancer related drugs, such as those for diabetes, to be repositioned for cancer treatment [89–93], our results also recommended drugs that are non-cancer related but may potentially be used for CRC patients. This is in addition to the majority of identified drugs that were originally designed for cancer-related therapy but mostly not in current use in CRC. For the drugs that were recommended to be repositioned, but so far not directly associated with CRC treatment, we determined that the abnormal genes of the given subgroups directly or indirectly interacted with the targets of these drugs, which had the highest aggregated scores in the subgroup's network. Additionally, wet lab-based *in vitro* experiments demonstrated that some of these drugs have a potential role in CRC treatment as a single or combinatorial therapy that may overcome resistance to other CRC drugs [69,74,94–99]. In our analysis of the top drugs for our subgroups of interest, enrichment analysis for these drugs was done based on their pharmacological class and pathways that are highly enriched as targets for the drugs. We found that topoisomerase inhibitors were the pharmacological class that had the highest number of drugs. Topoisomerase inhibitors affect the cell cycle and result in cell death. In addition to drugs that are already approved as CRC treatment in this class, others have demonstrated activity in metastatic colorectal cancer therapy [100]. Also, it can be used as a combined therapy to increase programmed cell death in CRC [101]. Regarding our analysis for the pathways, we found that the top pathways targeted by most top drugs are cancer-related pathways. Further improvements to our study can be made in the future. While we used published literature to validate our findings in addition to the randomized analysis that showed our recommended drugs were not selected by chance, further wet lab experiments to validate these findings are required prior to initiating clinical trials using these repurposed therapies.

6. Conclusion

In this study, we present a novel patient stratification and drug repositioning method using exploratory data mining and network analysis. The majority of currently available stratification methods are either not data-driven, in which biomarkers are used to stratify patients, or they are black box methods that are clinically challenging to implement due to lack of explainability. The proposed method uses patient genotypic and phenotypic data with a knowledge base that includes different biomedical entities to ensure the involvement of multiple biomedical aspects in our analysis. The subgrouping method identifies druggable homogeneous subgroups within a heterogeneous disease population. Colorectal cancer (CRC) was used as a case study. Our method found 130 CRC subgroups that were statistically significant. Together with medical experts, we focused on the subgroups that have microsatellite instability (MSI) status as a critical clinicopathologic variable. Seven subgroups were selected for further analysis to study their genotypic characteristics and the reasoning behind the suggested drugs for these subgroups. Our results were validated by a literature review demonstrating the relevance of these treatment options. Most of the suggested drugs for repositioning have known potential for cancer therapy and have the potential to be integrated in CRC treatment regimens. In the next phase of this research, we plan to validate these promising results at the bench in tumor cell lines *in vitro* and in vivo. In addition, drug resistance mechanisms and side effect profiles will be further studied in preparation for clinical translation.

CRediT authorship contribution statement

Zainab Al-Taie: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Danlu Liu:** Software, Writing - review & editing. **Jonathan Mitchem:** Conceptualization,

Writing - review & editing, Validation. **Christos Papageorgiou:** Validation. **Jussuf T. Kaifi:** Writing - review & editing, Validation. **Wesley C. Warren:** Validation. **Chi-Ren Shyu:** Conceptualization, Methodology, Validation, Writing - review & editing, Supervision, Project administration, Funding acquisition.

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 A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2021.103792.

References

[1] P. Deotarse, A. Jain, M. Baile, N. Kolhe, A. Kulkarni, Drug repositioning: a review, Int. J. Pharma. Res. Rev. 4 (2015) 51–58.
[2] S. Alaimo, A. Pulvirenti, Network-based drug repositioning: approaches, resources, and research directions, Methods Mol. Biol. (Clifton, NJ). 1903 (2019) 97–113.
[3] R.M. Plenge, E.M. Scolnick, D. Altshuler, Validating therapeutic targets through human genetics, Nat. Rev. Drug Discov. 12 (2013) 581–594.
[4] Z. Liu, H. Fang, K. Reagan, X. Xu, D.L. Mendrick, W. Slikker Jr, et al., In silico drug repositioning: what we need to know, Drug Discov Today. 18 (2013) 110–115.
[5] K. Park, A review of computational drug repurposing, Transl. Clin. Pharmacol. 27 (2019) 59–63.
[6] B. Readhead, J. Dudley, Translational bioinformatics approaches to drug development, Adv. Wound Care (New Rochelle). 2 (2013) 470–489.
[7] S. Keserci, E. Livingston, L. Wan, A.R. Pico, G. Chacko, Research synergy and drug development: Bright stars in neighboring constellations, Heliyon 3 (2017), e00442.
[8] H. Xue, J. Li, H. Xie, Y. Wang, Review of drug repositioning approaches and resources, Int. J. Biol. Sci. 14 (2018) 1232.
[9] R. Xu, Q. Wang, Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing, BMC Bioinf. 14 (2013) 181.
[10] M. Lotfi Shahreza, N. Ghadiri, S.R. Mousavi, J. Varshosaz, J.R. Green, Heter-LP: A heterogeneous label propagation algorithm and its application in drug repositioning, J. Biomed. Inform. 68 (2017) 167–183.
[11] G. Hu, P. Agarwal, Human disease-drug network based on genomic expression profiles, PLoS ONE 4 (2009), e6536.
[12] R. Xu, Q. Wang, PhenoPredict: A disease phenome-wide drug repositioning approach towards schizophrenia drug discovery, J. Biomed. Inform. 56 (2015) 348–355.
[13] M. Campillos, M. Kuhn, A.-C. Gavin, L.J. Jensen, P. Bork, Drug target identification using side-effect similarity, Science 321 (2008) 263–266.
[14] R. Xu, Q. Wang, A genomics-based systems approach towards drug repositioning for rheumatoid arthritis, BMC Genomics 17 (Suppl 7) (2016) 518.
[15] J. Lamb, E.D. Crawford, D. Peck, J.W. Modell, I.C. Blat, M.J. Wrobel, et al., The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease, Science 313 (2006) 1929–1935.
[16] J. Lamb, The Connectivity Map: a new tool for biomedical research, Nat. Rev. Cancer 7 (2007) 54.
[17] H. Liu, Y. Song, J. Guan, L. Luo, Z. Zhuang, Inferring new indications for approved drugs via random walk on drug-disease heterogenous networks, BMC Bioinf. 17 (2016) 539.
[18] Y. Wang, S. Chen, N. Deng, Y. Wang, Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data, PLoS ONE 8 (2013), e78518.
[19] Z. Tian, Z. Teng, S. Cheng, M. Guo, Computational drug repositioning using meta-path-based semantic network analysis, BMC Syst. Biol. 12 (2018) 134.
[20] B.K. Lee, K.H. Tiong, J.K. Chang, C.S. Liew, Z.A. Abdul Rahman, A.C. Tan, et al., DeSigN: connecting gene expression with therapeutics for drug repurposing and development, BMC Genomics 18 (2017) 934.
[21] F. Cheng, J. Zhao, M. Fooksa, Z. Zhao, A network-based drug repositioning infrastructure for precision cancer medicine through targeting significantly mutated genes in the human cancer genomes, J. Am. Med. Inform. Assoc.: JAMIA 23 (2016) 681–691.
[22] S.B. Metaphor, S.B.D. Nicholas, Time for one-person trials, Nature 520 (2015).
[23] D. Liu, W. Baskett, D. Beversdorf, C.-R. Shyu, Exploratory Data Mining for Subgroup Cohort Discoveries and Prioritization, IEEE J. Biomed. Health. Inf. (2019).
[24] D.S. Himmelstein, A. Lizee, C. Hessler, L. Brueggeman, S.L. Chen, D. Hadley, et al., Systematic integration of biomedical knowledge prioritizes drugs for repurposing, Elife. 6 (2017), e26726.
[25] L. Schneider, T. Kehl, K. Thedinga, N.L. Grammes, C. Backes, C. Mohr, et al., ClinOmicsTrailbc: a visual analytics tool for breast cancer treatment stratification, Bioinformatics (2019).
[26] Y. Chen, R. Xu, Drug repurposing for glioblastoma based on molecular subtypes, J. Biomed. Inform. 64 (2016) 131–138.
[27] B. Turanli, K. Karagoz, G. Bidkhori, R. Sinha, M.L. Gatza, M. Uhlen, et al., Multi-omic data interpretation to repurpose subtype specific drug candidates for breast cancer, Front. Genet. 10 (2019) 420.
[28] C. Zhou, X. Zhong, P. Gao, Z. Wu, J. Shi, Z. Guo, et al., Statin use and its potential therapeutic role in esophageal cancer: a systematic review and meta-analysis, Cancer Manage. Res. 11 (2019) 5655–5663.
[29] S. Gouravan, L.A. Meza-Zepeda, O. Myklebost, E.W. Stratford, E. Munthe, Preclinical evaluation of vemurafenib as therapy for BRAF(V600E) Mutated Sarcomas, Int. J. Mol. Sci. 19 (2018).
[30] L. Simon, V.P. Lavallee, M.E. Bordeleau, J. Krosl, I. Baccelli, G. Boucher, et al., Chemogenomic Landscape of RUNX1-mutated AML Reveals Importance of RUNX1 Allele Dosage in Genetics and Glucocorticoid Sensitivity, Clin. Cancer Res.: Off. J. Am. Assoc. Cancer Res. 23 (2017) 6969–6981.
[31] G.J. Yoshida, Emerging roles of Myc in stem cell biology and novel tumor therapies, J. Exp. Clin. Cancer Res. 37 (2018) 173.
[32] C. Nepal, C.J. O'Rourke, D. Oliveira, A. Taranta, S. Shema, P. Gautam, et al., Genomic perturbations reveal distinct regulatory networks in intrahepatic cholangiocarcinoma, Hepatology 68 (2018) 949–963.
[33] A.P. Lind, P.C. Anderson, Predicting drug activity against cancer cells by random forest models based on minimal genomic information and chemical properties, PLoS ONE 14 (2019), e0219774.
[34] V. Gligorijević, N. Malod-Dognin, N. Pržulj, Patient-specific data fusion for cancer stratification and personalised treatment, in: Biocomputing 2016: Proceedings of the Pacific Symposium: World Scientific, 2016, p. 321–332.
[35] A. Holzinger, A. Carrington, H. Müller, Measuring the Quality of Explanations: The System Causability Scale (SCS): Comparing Human and Machine Explanations, Kunstliche intelligenz. 34 (2020) 193–198.
[36] H. Hagras, Toward human-understandable, explainable AI, Computer. 51 (2018) 28–36.
[37] P.J. Slomka, R.J. Miller, I. Isgum, D. Dey, Application and Translation of Artificial Intelligence to Cardiovascular Imaging in Nuclear Medicine and Noncontrast CT, Semin. Nucl. Med. 50 (2020) 357–366.
[38] A.L. Harfouche, D.A. Jacobson, D. Kainer, J.C. Romero, A.H. Harfouche, G. Scarascia Mugnozza, et al., Accelerating Climate Resilient Plant Breeding by Applying Next-Generation Artificial Intelligence, Trends Biotechnol. 37 (2019) 1217–1235.
[39] A. Holzinger, M. Plass, M. Kickmeier-Rust, K. Holzinger, G.C. Crişan, C.-M. Pintea, et al., Interactive machine learning: experimental evidence for the human in the algorithmic loop, Appl. Intell. 49 (2019) 2401–2414.
[40] B.C. Wallace, K. Small, C.E. Brodley, J. Lau, T.A. Trikalinos, Deploying an interactive machine learning system in an evidence-based practice center: abstrackr, in: Proceedings of the 2nd ACM SIGHIT international health informatics symposium, 2012, p. 819–824.
[41] S. Teso, K. Kersting, Explanatory interactive machine learning, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, p. 239–245.
[42] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:170208608. 2017.
[43] G. Dong, J. Li, Efficient mining of emerging patterns: Discovering trends and differences, in: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 1999, p. 43–52.
[44] P. Pudil, J. Novovičová, J. Kittler, Floating search methods in feature selection, Pattern Recogn. Lett. 15 (1994) 1119–1125.
[45] R. Agarwal, R. Srikant, Fast algorithms for mining association rules. Proc of the 20th VLDB Conference, 1994, p. 487–499.
[46] A. Jain, K. Nandakumar, A. Ross, Score normalization in multimodal biometric systems, Pattern Recogn. 38 (2005) 2270–2285.
[47] C. Wu, R.C. Gudivada, B.J. Aronow, A.G. Jegga, Computational drug repositioning through heterogeneous network clustering, BMC Syst. Biol. 7 (Suppl 5) (2013) S6.
[48] L. Yu, S. Yao, L. Gao, Y. Zha, Conserved Disease Modules Extracted From Multilayer Heterogeneous Disease and Gene Networks for Understanding Disease Mechanisms and Predicting Disease Treatments, Front. Genet. 9 (2018) 745.
[49] X. Qi, M. Shen, P. Fan, X. Guo, T. Wang, N. Feng, et al., The Performance of Gene Expression Signature-Guided Drug-Disease Association in Different Categories of Drugs and Diseases, Molecules (Basel, Switzerland) 25 (2020).
[50] T. Lee, Y. Yoon, Drug repositioning using drug-disease vectors based on an integrated network, BMC Bioinf. 19 (2018) 446.
[51] M. Iwata, L. Hirose, H. Kohara, J. Liao, R. Sawada, S. Akiyoshi, et al., Pathway-Based Drug Repositioning for Cancers: Computational Prediction and Experimental Validation, J. Med. Chem. 61 (2018) 9583–9595.
[52] Z. Wu, Y. Wang, L. Chen, Drug repositioning framework by incorporating functional information, IET Syst. Biol. 7 (2013) 188–194.
[53] M. Kissa, G. Tsatsaronis, M. Schroeder, Prediction of drug gene associations via ontological profile similarity with application to drug repositioning, Methods 74 (2015) 71–82.
[54] Y. Zheng, H. Peng, X. Zhang, Z. Zhao, X. Gao, J. Li, Old drug repositioning and new drug discovery through similarity learning from drug-target joint feature spaces, BMC Bioinf. 20 (2019) 605.
[55] Y. Taguchi, T. Turki, Universal Nature of Drug Treatment Responses in Drug-Tissue-Wide Model-Animal Experiments Using Tensor Decomposition-Based Unsupervised Feature Extraction, Front. Genet. 11 (2020) 695.

[56] Z. Liu, F. Guo, J. Gu, Y. Wang, Y. Li, D. Wang, et al., Similarity-based prediction for Anatomical Therapeutic Chemical classification of drugs by integrating multiple data sources, Bioinformatics 31 (2015) 1788–1795.

[57] P.N. Hameed, K. Verspoor, S. Kusljic, S. Halgamuge, A two-tiered unsupervised clustering approach for drug repositioning through heterogeneous data integration, BMC Bioinf. 19 (2018) 129.

[58] Y. Sun, P.N. Hameed, K. Verspoor, S. Halgamuge, A physarum-inspired prize-collecting steiner tree approach to identify subnetworks for drug repositioning, BMC Syst. Biol. 10 (2016) 128.

[59] S.D. Markowitz, M.M. Bertagnolli, Molecular basis of colorectal cancer, N. Engl. J. Med. 361 (2009) 2449–2460.

[60] B. Vogelstein, E.R. Fearon, S.R. Hamilton, S.E. Kern, A.C. Preisinger, M. Leppert, et al., Genetic alterations during colorectal-tumor development, N. Engl. J. Med. 319 (1988) 525–532.

[61] S. Hasan, P. Renz, R.E. Wegner, G. Finley, M. Raj, D. Monga, et al., Microsatellite instability (MSI) as an independent predictor of pathologic complete response (PCR) in locally advanced rectal cancer: a National Cancer Database (NCDB) Analysis, Ann. Surg. 271 (2020) 716–723.

[62] T. André, A. De Gramont, D. Vernerey, B. Chibaudel, F. Bonnetain, A. Tijeras-Raballand, et al., Adjuvant fluorouracil, leucovorin, and oxaliplatin in stage II to III colon cancer: updated 10-year survival and outcomes according to BRAF mutation and mismatch repair status of the MOSAIC study, J. Clin. Oncol. 33 (2015) 4176–4187.

[63] D.T. Le, J.N. Uram, H. Wang, B.R. Bartlett, H. Kemberling, A.D. Eyring, et al., PD-1 blockade in tumors with mismatch-repair deficiency, N. Engl. J. Med. 372 (2015) 2509–2520.

[64] O. Murcia, M. Juárez, M. Rodríguez-Soler, E. Hernández-Illán, M. Giner-Calabuig, M. Alustiza, et al., Colorectal cancer molecular classification using BRAF, KRAS, microsatellite instability and CIMP status: Prognostic implications and response to chemotherapy, PLoS ONE 13 (2018), e0203051.

[65] F. Loupakis, D. Yang, L. Yau, S. Feng, C. Cremolini, W. Zhang, et al., Primary tumor location as a prognostic factor in metastatic colorectal cancer, J. Natl Cancer Inst. 107 (2015).

[66] G. Nyamundanda, E. Fontana, A. Sadanandam, Is the tumour microenvironment a critical prognostic factor in early-stage colorectal cancer? Ann. Oncol.: Off. J. Eur. Soc. Med. Oncol. 30 (2019) 1538–1540.

[67] Y. Yang, G. Wang, J. He, S. Ren, F. Wu, J. Zhang, et al., Gender differences in colorectal cancer survival: A meta-analysis, Int. J. Cancer 141 (2017) 1942–1949.

[68] R. Dienstmann, G. Villacampa, A. Sveen, M.J. Mason, D. Niedzwiecki, A. Nesbakken, et al., Relative contribution of clinicopathological variables, genomic markers, transcriptomic subtyping and microenvironment features for outcome prediction in stage II/III colorectal cancer, Ann. Oncol.: Off. J. Eur. Soc. Med. Oncol. 30 (2019) 1622–1629.

[69] C. Kishore, S. Sundaram, D. Karunagaran, Vitamin K3 (menadione) suppresses epithelial-mesenchymal-transition and Wnt signaling pathway in human colorectal cancer cells, Chem. Biol. Interact. 309 (2019), 108725.

[70] M.F. Hegazy, M. Fukaya, M. Dawood, G. Yan, A. Klinger, E. Fleischer, et al., Vitamin K(3) thio-derivative: a novel specific apoptotic inducer in the doxorubicin-sensitive and -resistant cancer cells, Invest. New Drugs 38 (2020) 650–661.

[71] Y. Nakamura, T. Yamaguchi, Stereoselective metabolism of 2-phenylpropionic acid in rat. I. In vitro studies on the stereoselective isomerization and glucuronidation of 2-phenylpropionic acid, Drug Metabolism Disposition: Biol. Fate Chem. 15 (1987) 529–534.

[72] F. Du, X. Li, W. Feng, C. Qiao, J. Chen, M. Jiang, et al., SOX13 promotes colorectal cancer metastasis by transactivating SNAI2 and c-MET, Oncogene 39 (2020) 3522–3540.

[73] R. Graves-Deal, G. Bogatcheva, S. Rehman, Y. Lu, J.N. Higginbotham, B. Singh, Broad-spectrum receptor tyrosine kinase inhibitors overcome de novo and acquired modes of resistance to EGFR-targeted therapies in colorectal cancer, Oncotarget. 10 (2019) 1320–1333.

[74] K.C. Cuneo, R.K. Mehta, H. Kurapati, D.G. Thomas, T.S. Lawrence, M.K. Nyati, Enhancing the Radiation Response in KRAS Mutant Colorectal Cancers Using the c-Met Inhibitor Crizotinib, Transl. Oncol. 12 (2019) 209–216.

[75] O. Sordet, Q.A. Khan, K.W. Kohn, Y. Pommier, Apoptosis induced by topoisomerase inhibitors, Curr. Med. Chem. Anticancer Agents 3 (2003) 271–290.

[76] A. Dehshahri, M. Ashrafizadeh, E. Ghasemipour Afshar, A. Pardakhty, A. Mandegary, R. Mohammadinejad, et al., Topoisomerase inhibitors: Pharmacology and emerging nanoscale delivery systems, Pharmacol. Res. 151 (2020), 104551.

[77] S. Jacob, M. Aguado, D. Fallik, F. Praz, The role of the DNA mismatch repair system in the cytotoxicity of the topoisomerase inhibitors camptothecin and etoposide to human colorectal cancer cells, Cancer Res. 61 (2001) 6555–6562.

[78] S. Stintzing, H.J. Lenz, Protein kinase inhibitors in metastatic colorectal cancer. Let's pick patients, tumors, and kinase inhibitors to piece the puzzle together!, Expert Opin. Pharmacother. 14 (2013) 2203–2220.

[79] S.B. Nygård, I.J. Christensen, D.H. Smith, S.L. Nielsen, N.F. Jensen, H.J. Nielsen, et al., Underpinning the repurposing of anthracyclines towards colorectal cancer: assessment of topoisomerase II alpha gene copy number alterations in colorectal cancer, Scand. J. Gastroenterol. 48 (2013) 1436–1443.

[80] L.S. Tarpgaard, C. Qvortrup, S.B. Nygård, S.L. Nielsen, D.R. Andersen, N. F. Jensen, et al., A phase II study of Epirubicin in oxaliplatin-resistant patients with metastatic colorectal cancer and TOP2A gene amplification, BMC cancer. 16 (2016) 1–5.

[81] J.I. Lai, Y.J. Tseng, M.H. Chen, C.F. Huang, P.M. Chang, Clinical Perspective of FDA Approved Drugs With P-Glycoprotein Inhibition Activities for Potential Cancer Therapeutics, Front. Oncol. 10 (2020), 561936.

[82] E.Y. Chen, C.M. Tan, Y. Kou, Q. Duan, Z. Wang, G.V. Meirelles, et al., Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool, BMC Bioinf. 14 (2013) 128.

[83] M.V. Kuleshov, M.R. Jones, A.D. Rouillard, N.F. Fernandez, Q. Duan, Z. Wang, et al., Enrichr: a comprehensive gene set enrichment analysis web server 2016 update, Nucleic Acids Res. 44 (2016) W90–W97.

[84] Q. Fan, B. Liu, Identification of the anticancer effects of a novel proteasome inhibitor, ixazomib, on colorectal cancer using a combined method of microarray and bioinformatics analysis, OncoTargets Therapy 10 (2017) 3591–3606.

[85] L.C. Ye, T. Chen, D.X. Zhu, S.X. Lv, J.J. Qiu, J. Xu, et al., Downregulated long non-coding RNA CLMAT3 promotes the proliferation of colorectal cancer cells by targeting regulators of the cell cycle pathway, Oncotarget. 7 (2016) 58931–58938.

[86] J. Tan, L. Zhuang, H.S. Leong, N.G. Iyer, E.T. Liu, Q. Yu, Pharmacologic modulation of glycogen synthase kinase-3beta promotes p53-dependent apoptosis through a direct Bax-mediated mitochondrial pathway in colorectal cancer cells, Cancer Res. 65 (2005) 9012–9020.

[87] C. Richardson, S. Zhang, L.J. Hernandez Borrero, W.S. El-Deiry, Small-molecule CB002 restores p53 pathway signaling and represses colorectal cancer cell growth, Cell Cycle 16 (2017) 1719–1725.

[88] S. Attoub, C. Rivat, S. Rodrigues, S. Van Bocxlaer, M. Bedin, E. Bruyneel, et al., The c-kit tyrosine kinase inhibitor STI571 for colorectal cancer therapy, Cancer Res. 62 (2002) 4879–4883.

[89] B.J. Quinn, H. Kitagawa, R.M. Memmott, J.J. Gills, P.A. Dennis, Repositioning metformin for cancer prevention and treatment, Trends Endocrinol. Metab. 24 (2013) 469–480.

[90] N. Saini, X. Yang, Metformin as an anti-cancer agent: actions and mechanisms targeting cancer stem cells, Acta Biochim. Biophy. Sin. 50 (2018) 133–143.

[91] J. Wojciechowska, W. Krajewski, M. Bolanowski, T. Kręcicki, T. Zatoński, Diabetes and Cancer: a Review of Current Knowledge, Exp. Clin. Endocrinol. Diabetes: Off. J. German Soc. Endocrinol. [and] German Diabetes Assoc. 124 (2016) 263–275.

[92] G.R. Jones, M.P. Molloy, Metformin, Microbiome and Protection Against Colorectal Cancer, Digestive Diseases and Sciences 2020.

[93] A. Gadducci, N. Biglia, R. Tana, S. Cosio, M. Gallo, Metformin use and gynecological cancers: A novel treatment option emerging from drug repositioning, Crit. Rev. Oncol./Hematol. 105 (2016) 73–83.

[94] I. De Pauw, F. Lardon, J. Van den Bossche, H. Baysal, P. Pauwels, M. Peeters, et al., Overcoming Intrinsic and Acquired Cetuximab Resistance in RAS Wild-Type Colorectal Cancer: An In Vitro Study on the Expression of HER Receptors and the Potential of Afatinib, Cancers. 11 (2019).

[95] M. Yang, X. Fang, J. Li, D. Xu, Q. Xiao, S. Yu, et al., Afatinib treatment for her-2 amplified metastatic colorectal cancer based on patient-derived xenograft models and next generation sequencing, Cancer Biol. Ther. 20 (2019) 391–396.

[96] E.F. Dunn, M. Iida, R.A. Myers, D.A. Campbell, K.A. Hintz, E.A. Armstrong, et al., Dasatinib sensitizes KRAS mutant colorectal tumors to cetuximab, Oncogene 30 (2011) 561–574.

[97] G. Rao, I.K. Kim, F. Conforti, J. Liu, Y.W. Zhang, G. Giaccone, Dasatinib sensitises KRAS-mutant cancer cells to mitogen-activated protein kinase kinase inhibitor via inhibition of TAZ activity, Eur. J. Cancer (Oxford, England 2018 (99) (1990) 37–48.

[98] C.B. Williams, C. McMahon, S.M. Ali, M. Abramovitz, K.A. Williams, J. Klein, et al., A metastatic colon adenocarcinoma harboring BRAF V600E has a durable major response to dabrafenib/trametinib and chemotherapy, OncoTargets Therapy 8 (2015) 3561–3564.

[99] S.W. Leung, C.J. Chou, T.C. Huang, P.M. Yang, An Integrated Bioinformatics Analysis Repurposes an Antihelminthic Drug Niclosamide for Treating HMGA2-Overexpressing Human Colorectal Cancer, Cancers. 11 (2019).

[100] W.H. Isacoff, K. Borud, Chemotherapy for the treatment of patients with metastatic colorectal cancer: an overview, World J. Surg. 21 (1997) 748–762.

[101] A. Pawlak, E. Ziolo, A. Fiedorowicz, K. Fidyt, L. Strzadala, W. Kalas, Long-lasting reduction in clonogenic potential of colorectal cancer cells by sequential treatments with 5-azanucleosides and topoisomerase inhibitors, BMC Cancer 16 (2016) 893.