Genetics and Comorbidity network analysis of Colorectal cancer

Branimir Ljubic (✉ branimir.ljubic@temple.edu)

Temple University <https://orcid.org/0000-0002-3287-3741>

Martin Pavlovski

Temple University

Jumanah Alshehri

Temple University

Vladimir Bajic

King Abdullah University of Science and Technology

Christophe Van Neste

King Abdullah University of Science and Technology

Zoran Obradovic

Temple University

Shoumik Roychoudhury

Temple University

Research article

Keywords: Colorectal cancer, Comorbidities, Genes, Treatment procedures, Social networks, Biomedical informatics

Posted Date: August 30th, 2019

DOI: <https://doi.org/10.21203/rs.2.13187/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Informatics in Medicine Unlocked on January 1st, 2020. See the published version at <https://doi.org/10.1016/j.imu.2020.100492>.

TITLE:

Genetics and Comorbidity Network Analysis of Colorectal Cancer

Authors:

Branimir Ljubic ¹, Martin Pavlovski ¹, Jumanah Alshehri ¹, Shoumik Roychoudhury ¹, Vladimir Bajic ², Christophe Van Neste ², Zoran Obradovic ¹

1. Temple University, Center for Data Analytics and Biomedical Informatics (DABI),
Philadelphia, PA, U.S.A.
2. King Abdullah University of Science and Technology, Computational Bioscience
Research Center, Thuwal, Kingdom of Saudi Arabia

Corresponding author:

Zoran Obradovic

Temple University, Center for Data Analytics and Biomedical Informatics, Philadelphia, PA,
U.S.A.

Mailing address:

Temple University, SERC 035-02
Center for Data Analytics and Biomedical Informatics
1925 N 12th St, Philadelphia, PA 19122

Email:

zoran.obradovic@temple.edu

Phone number:

+215-204-6265

ABSTRACT

Background

Colorectal cancer (CRC) is the third most common cancer in the United States, and the second leading cause of cancer death. Comorbidity network analyses of CRC can help understanding of the illness progression. About 10% - 30% of patients have a family history of CRC that suggests a hereditary contribution, including pathogenic variants of genes. The goal was to identify comorbidities associated with CRC and to discover pathogenic variants of genes that could be connected to CRC and comorbidity diseases.

Results

A novel model is developed based on comorbidity networks, to study progression of CRC. The model was developed on the HCUP, SID California inpatient database for a period of 9 years (2003-2011). Comorbidity networks, and Venn diagrams show probabilities of occurrence of comorbidities associated with different stages of CRC. Ranked lists of comorbidities (more than 5,800) were utilized for text mining of PubMed and expert curated sources (DisGeNet), to identify genes associated with CRC. Associations between 1,940 different genes and CRC were extracted from PubMed. 150 different genes are associated with CRC in DisGeNet. 96 genes are present in both sources (PubMed and curated sources). The most mentioned gene associated with CRC were: TP53 (241 abstracts in PubMed), APC (115), and KRAS (106). All 3 genes had DisGeNet scores of 0.5. Two more genes (MLH1 – 98 abstracts and TGFBR2 - 18 abstracts) had the same DisGeNet scores. PPARG gene (43 abstracts) had DisGeNet score of 0.6.

Genes described on the cancer.gov website were found on PubMed too. MUTYH was described in association with CRC in 7 abstracts, MSH6 in 25 abstracts, PMS2 in 11, EPCAM in 10, POLD1 in 4, BMPR1A in 3, SMAD4 in 15, PTEN in 25 and STK11 in 8 PubMed abstracts.

Lists of associations of the six most common genes and comorbidities of CRC were also created.

Conclusions

Results of the study could be considered as a novelty contributing to better understanding of risk factors (genes, comorbidities) for the development of CRC. Genetic findings could lead to the improvement of guidelines for genetic testing of patients that are at increased risk to develop CRC.

Keywords:

Colorectal cancer, Comorbidities, Genes, Treatment procedures, Social networks, Biomedical informatics

Background

There are more than 140,000 new cases and 51,000 deaths from colorectal cancer (CRC) in the U.S. each year.[1,2] CRC is the third most common cancer in the U.S., and the second leading cause of cancer death. When discovered early, it is highly treatable. Surgical treatment followed by chemotherapy is highly effective.[2,3] The stage (extent of spread) of a CRC depends on how deeply it grows into the wall and if it has spread outside the colon or rectum. Adenocarcinomas of the colon make up about 96% of CRC.[1,4,5]

Comorbidity network analyses of cancers can help understanding of the illness progression.[6,7]

Many applications of social networks, machine learning, and biomedical informatics methods related to CRC have been published.[8-10]

Numerous factors suggest a genetic contribution to CRC. Some of these factors are a family history of CRC or polyps, multiple primary cancers in a patient with CRC, etc.[11] Several genes associated with CRC risk have been identified. “Hereditary CRC has two forms: (1) polyposis, caused by variants in the APC gene; and MUTYH-associated polyposis, caused by variants in the MUTYH gene and (2) Lynch syndrome, caused by variants in DNA MMR genes (MLH1, MSH2, MSH6 and PMS2) and EPCAM. Other CRC syndromes and their associated genes include oligopolyposis (POLE, POLD1), NTHL1, juvenile polyposis syndrome (BMPR1A, SMAD4), Cowden syndrome (PTEN), and Peutz-Jeghers syndrome (STK11)”.[11] About 10% to 30% of patients have a family history of CRC that suggests a hereditary contribution, including pathogenic variants of genes.[12]

The first objective of our research was to utilize a network science, to discover the most common comorbidities and procedures and their significance and probabilities of development in different

stages of CRC progression. It has been known that adverse effects of comorbidities could exacerbate and worsen the course of the cancer.

Comorbidities and procedures associated with CRC registered in the Healthcare Cost and Utilization Project (HCUP) databases have been discovered and analyzed. In order to contribute to better understanding of the cancer disease, scale-free comorbidity networks have been created based on Pierson's correlation and Relative Risk (RR) among the most common comorbidities. Centrality measures applicable to these networks were calculated to determine the significance of the most common comorbidities.

We used discovered comorbidities of CRC to find associations between genes, CRC, and comorbidity diseases. The text mining tool, BeFree was utilized to extract relationships between CRC and genes from PubMed.[13] BeFree consists of the Biomedical Named Entity Recognition (BioNER) module based on dictionaries using fuzzy and pattern matching methods to find and uniquely identify entity mentions in the literature, and a module for Relation Extraction (RE) based on Support Vector Machine (SVM).[13] Relations between biomedical entities are recognized by exploiting syntactic and semantic information from the text.

Findings from BeFree text mining were compared with results from expert curated sources of known genes. We used DisGeNET, which is one of the largest collections of genes involved in human diseases. DisGeNET integrates data from expert curated repositories, the GWAS catalogue, animal models and scientific publications.[14]

The second objective of our study was to discover relations between CRC and comorbidities of CRC with genes in abstracts indexed in the PubMed database. The goal was to identify pathogenic variants of genes that could potentially be connected to CRC and comorbidity diseases. Results of text mining using BeFree, were compared with curated sources of known genes. We checked how

many genes related to CRC and its comorbidities from searches with BeFree overlap results from searches of curated sources. Discovered additional genes not registered in expert curated sources, and associated with CRC and their relations with other diseases could be further evaluated in cancer genetic studies as the potential risk factor for development of CRC.

In May 2015, the Cancer and Aging Research Group in collaboration with the National Cancer Institute and National Institute on Aging, analyzed the relevance of comorbidities in cancer.[15]

In a study of Medicare beneficiaries, over two-thirds had two or more medical conditions and nearly 25% of participants had four or more conditions. Most clinical practice guidelines are single-disease focused, limiting their application in patients with multiple comorbid illnesses.[16] Most studies have found that patients with cancer and comorbidities have worse survival than those without comorbidity.[17]

Hidalgo and Barabasi described the use of networks to integrate different genetic, proteomic, and metabolic datasets as a viable path toward elucidating the origins of specific diseases.[6] The research presented evidence that the structure of the Phenotypic Disease Network (PDN) is relevant to the understanding of illness progression. Patient age and comorbidity affect survival of CRC.[18] Another study evaluated trends in the prevalence and impact of comorbidity on CRC survival.[19]

The genetic basis of hereditary CRC and survival of carriers of genetic mutations has been described. [20,21] The American College of Medical Genetics and Genomics has published guidelines for evaluating patients with suspected CRC susceptibility syndromes, with the goal to identify individuals whose clinical findings require referral for genetics consultation.[22]

CRC has a high genetic heterogeneity, and it has been shown that some rare mutations in CRC are actually quite common and may be involved in the occurrence of other types of cancer as well.[23]

It is becoming increasingly clear that there are genetic subsets of CRC. There has been a significant expansion of the molecular and genetic understanding of colonic carcinogenesis in the last 30 years.[24]

Methods

We performed analyses of comorbidities and procedures associated with CRC on the HCUP - State Inpatient Databases (SID), which includes the largest collection of longitudinal hospital care data in the U.S.[25] We analyzed data from the HCUP - SID California inpatient database for a period of 9 years (2003-2011).

A novel network science-based method to study progression of CRC was developed. Our comprehensive approach comprises comorbidity networks, ranked lists and Venn diagrams with probabilities of occurrence of comorbidities and procedures associated with CRC. Downloaded data were analyzed separately in early and late stages of CRC and separately for males and females. Fundamentally, we considered all patients with CRC without diagnosed metastases as early stages of the cancer. Patients who had diagnosed metastases were included in advanced stages of the cancer.

Rank lists of procedures (by prevalence) performed in patients with CRC and comorbidities of CRC, have been created for early and advanced stages of the cancer, for males and females. Venn diagrams visualize evolvement of comorbidities and procedures as the cancer progresses and show probabilities of occurrences of each of the comorbidities and procedures.

Two types of comorbidity networks were constructed, one based on Pearson's correlations and the other based on RR. In order to quantify the strength of relations among comorbidities a "distance" between two diseases was measured. Two comorbidity measures were used to quantify the

distance between two diseases: RR and Pearson's correlation.[6] The RR of observing a pair of diseases i and j affecting the same patient is given by the following formula:

$$RR_{ij} = \frac{C_{ij}N}{P_i P_{ij}} \quad (1)$$

where C_{ij} is the number of patients affected by both diseases, N is the total number of patients in the population and P_i and P_j are the prevalence of diseases i and j .

Pearson's-correlation (ϕ -correlation) is computed as:

$$\phi_{ij} = \frac{C_{ij}N - P_i P_{ij}}{\sqrt{P_i P_j (N - P_i)(N - P_j)}} \quad (2)$$

Correlation and RR increase with the number of patients affected by both diseases. The problem with RR is that it overestimates relationships involving rare diseases and underestimates the comorbidity between highly prevalent illnesses. Pearson's correlation accurately discriminates comorbidities between pairs of diseases of similar prevalence but underestimates the comorbidity between rare and common diseases. In both models, the top 100 comorbidities from the rank lists were used to construct adjacency matrix, $A[ij]$ which encodes whether and how a pair of nodes is connected. We constructed a signed correlation network, which was created from the adjacency matrix and performed power transformation (normalization) of correlations.[26] For normalization, we used a signed network normalization $|(corMatrix+1)/2|^\beta$ for a 100 x 100 adjacency matrix. By raising the absolute value of the correlation to a power $\beta \geq 1$ (soft

thresholding), the correlation network emphasizes high correlations at the expense of low correlations. We tested β -values between 1 and 10 and chose $\beta = 2$ for Pearson's correlation and $\beta = 5$ for RR. We used the correlation matrix after the transformation (normalization) as adjacent matrices to plot the network. The cutoff for edges, to be plotted on the network was set to some reasonable number. This method was applied to construct all 8 networks. Centrality measures (degrees of nodes and betweenness) were calculated to analyze the significance of each of the comorbidities.

Further, we leveraged and exploited lists of ranked comorbidities of CRC for text mining of PubMed, in order to identify genes associated with CRC and comorbidities. The BeFree text mining system was utilized to extract associations between CRC, ranked comorbidity diseases and genes by using morpho-syntactic information of the text. Further we extracted relationships between genes and diseases from expert curated sources (DisGeNET).

We compared a proportion of gene-comorbidity disease associations discovered using BeFree with data collected in expert-curated databases (DisGeNET). Findings were also compared with relationships between genes and CRC published on the cancer.gov website. Genes most often described in associations with CRC were further matched with comorbidities from the ranking list. The number of abstracts in PubMed where these associations were described was extracted and counted. Lists of associations of the six most common genes and comorbidities of CRC were created in order to show the number of abstracts where these associations were detected.

Results

The comorbidity network research was conducted on the HCUP, SID California database for the period from 2003 to 2011. The total number of females with early stages of CRC was 31,503, and

males was 30,870. We found 2,019 different procedures in females and 1,930 procedures in males. These procedures include all procedures for treatment of all comorbidities that patients with CRC developed. The most common procedures performed in patients with CRC are different than what we can expect to find in the general population. The majority of the top 30 procedures are related to different aspects of CRC treatment, with the addition of several procedures related to cardiovascular conditions.

The total number of patients with late stage CRC was: females – 26,906 and males – 27,454. We identified 1,953 different procedures in females and 1,988 procedures in males in advanced stages of CRC. With the exception of few procedures, specific to cardiovascular conditions, more than 90% of procedures are specific to the treatment of CRC. An interesting discovery is that the top 30 procedures in females and males were almost identical except some differences in the ranking order of procedures. The top procedure in early and advanced stages in both genders was transfusion of packed cells.

Venn diagrams in figure 1 present comparative changes of treatment procedures with the cancer progression from early to advanced stages. Procedures presented in Venn diagrams include all procedures for any possible comorbidity condition that patients with CRC developed. Venn diagrams also show probabilities of each of top 30 procedures in early and late stages.

Inspection of additional top 300 procedures show that laparoscopic and endoscopic procedures had lower ranks than open invasive procedures, which can be considered for reversal in the future.

Also, ultrasound of abdomen has lower probability of being performed than CT of abdomen.

Figure 1. Comparative Venn diagrams of the top 30 procedures in females and males.

Left: Ranked procedures in females performed only in early stages (top), procedures performed in both stages (middle), and procedures performed only in advanced stages (bottom). Right: Ranked procedures in males performed only in early stages (top), procedures performed in both stages (middle), and procedures performed only in advanced stages (bottom). Probabilities for each procedure are shown. The first number in both stages corresponds to probabilities of procedures in early stages, and the second number shows probabilities of procedures in advanced stages of colorectal cancer.

Further, we conducted analyses of comorbidities in patients with early stages of CRC. We identified 5,786 different comorbidity conditions in females and 5,607 in males. Among the top 30 comorbidities, we notice specific diseases for patients with history of CRC (benign neoplasms of colon, diverticulosis, paralytic ileus, history of malignant neoplasm of colon...) which is different than in the general population. Not surprisingly, the top comorbidity was essential

hypertension, followed by metabolic conditions like hyperlipidemia, hypercholesterolemia, diabetes mellitus type II and cardiovascular conditions.

We discovered 5,609 different comorbidity conditions in females and 5,427 in males in advanced stages of CRC. There are few conditions specific to patients with CRC, but most of conditions on this list are similar to findings in the general population.

Venn diagrams in figure 2 show comparative evolvement of ranked comorbidities, along the progression of CRC from early toward advanced stages. The left side of the diagram shows the most often comorbidities in females and the right side the most often comorbidities in males. The diagram also shows probabilities that particular comorbidities will occur in early and late stages of the cancer. An important finding is that most of the top 30 comorbidities appear in both stages of the cancer.

generated based on these 8 matrices. In order to evaluate significance of each of the comorbidities, centrality measures, node degrees and betweenness, were calculated for all 8 networks.

In figure 3, we present two networks for early stages of CRC in females. Both networks considered the top 100 comorbidities. We calculated ϕ -correlation and RR and constructed 2 networks separately for each of these metrics.

Figure 3. Networks of comorbidities in early stages of colorectal cancer in females

Comorbidity networks, based on a) Pearson's correlation and b) RR in females. Nodes represent top 100 comorbidities (ICD9 codes). Correlations greater than 0.32 ($\beta = 2$) and RR greater than 5.99 ($\beta = 5$) were applied for construction of edges.

Considering ϕ -correlation, the following comorbidity conditions had the highest node degrees in network 3a: urinary tract infection(UTI), congestive heart failure(CHF), and coronary artery disease(CAD). Comorbidities that had the highest betweenness in network 3a were: UTI, CHF, CAD, and colostomy status.

Based on RR measures, comorbidities with the highest node degrees in network 3b were: anemia, obstructive chronic bronchitis, and hypoxemia. The following conditions had the highest betweenness in network 3b: postoperative infection, colostomy status, gastro-duodenitis, and chest pain.

Figure 4 shows two networks (ϕ -correlations and RR) for late stages of colorectal cancer in females. Both networks contain the top 100 comorbidities. Nodes represent comorbidities and links were created from the correlation of each pair of comorbidities.

Figure 4. Networks of comorbidities in advanced stages of colorectal cancer in females

Comorbidity networks, based on a) Pearson's correlation and b) RR in females. Nodes represent top 100 comorbidities (ICD9 codes). Correlations greater than 0.30 ($\beta = 2$) and RR greater than 8.99 ($\beta = 5$) were applied for construction of edges.

The highest node degrees in network 4a (ϕ -correlation) were: UTI, anemia, CHF, hypotassemia, and pneumonia and the highest betweenness: UTI, anemia, dehydration, and history of irradiation.

Considering RR measure, the highest node degrees in network 4b were: Chronic kidney disease (CKD), pressure ulcer- lower back, hyperkalemia, and sepsis. The highest betweenness in network 4b were: CKD, obstructive chronic bronchitis, history of antineoplastic chemotherapy, pressure ulcer-lower back and hypoxemia.

Next, we present two networks for early stages of colorectal cancer in males (figure 5). Both networks considered the top 100 comorbidities. We used the same metrics as in females, ϕ -correlations Fig5a. and RR – Fig5b.

Figure 5. Networks of comorbidities in early stages of colorectal cancer in males.

Comorbidity networks, based on a) Pearson's correlation and b) RR in males. Nodes represent top 100 comorbidities (ICD9 codes). Correlations greater than 0.32 ($\beta = 2$) and RR greater than 8.99 ($\beta = 5$) were applied for construction of edges.

Comorbidities with the highest node degrees in network 5a. (ϕ -correlation) were: CHF, acute kidney failure (AKF), anemia, and UTI, and with the highest betweenness: UTI, anemia, CHF, AKF, and colostomy status. The highest node degrees in network 5b, based on RR were: anemia,

hyperkalemia, rehabilitation procedure, and obstructive chronic bronchitis. The highest betweenness in network 5b were: colostomy status, anemia, postoperative infection, and diabetes type II.

We created two networks for advanced stages of CRC in males, considering the top 100 comorbidities (figure 6). We used the same metrics as in females, ϕ -correlation (Fig 6a) and RR (Fig 6b).

Figure 6. Networks of comorbidities in advanced stages of colorectal cancer in males.

Comorbidity networks, based on a) Pearson's correlation and b) RR. Nodes represent top 100 comorbidities (ICD9 codes). Correlations greater than 0.30 ($\beta = 2$) and RR greater than 8.99 ($\beta = 5$) were used for edges.

Conditions with the highest node degrees in network 6a. (ϕ -correlation) were: anemia, CHF, hypertensive chronic kidney disease, AKF, CKD, and UTI. Conditions with the highest betweenness in network 6a were: anemia, dehydration, CHF, and AKF.

408 Based on RR, comorbidities with the highest node degrees in network 6b were: Hypertensive
409 chronic kidney disease, stage V, primary cardiomyopathies and CKD. Comorbidities that had the
410 highest betweenness in network 6b were: Dysphagia, hypertensive CKD, persistent mental
411 disorders, and primary cardiomyopathies.

412 Ranked lists of comorbidities of CRC were used for text mining of PubMed and expert curated
413 sources, specifically DisGeNet, which is a comprehensive platform integrating information on
414 human disease-associated genes and variants. The BeFree text mining system was applied for the
415 extraction of data from PubMed related to associations between genes and CRC as well as between
416 genes and comorbidities of CRC. We extracted a full set of genes - disease relations from PubMed
417 abstracts and curated sources (DisGeNet).

418 MEDLINE (PubMed) currently indexes more than 190,000 publications on colorectal cancer, thus
419 it is a good resource to gather information on genetic determinants of this illness.

420 The Befree data mining system extracted 1,940 different genes from PubMed, associated with
421 CRC. We found 150 different genes associated with CRC in expert curated sources (DisGeNet).
422 Ninety six genes are present in both sources (PubMed and curated sources). Most of gene – CRC
423 associations in PubMed were described in only one abstract. (1160 – 60%). Vast majority of these
424 relations had low DisGeNet score – 0.01. The DisGeNET score for gene – disease associations
425 takes into account the number and type of sources (level of curation), and the number of
426 publications supporting the association. The scores range from 0 to 1.

427 Genes – CRC relationships were described 101 times in 10 or more abstracts. There are 3 genes
428 associated with CRC mentioned in more than 100 abstracts. The most mentioned genes associated
429 with CRC were: TP53 (241 abstracts in PubMed), APC (115 abstracts) and KRAS (106). All 3
430 genes had DisGeNet scores of 0.5. There were 2 more genes (MLH1 – 98 abstracts and TGFBR2

in only 18 abstracts) with the same DisGeNet scores of 0.5. One gene (PPARG) had DisGeNet score of 0.6. This gene was mentioned 43 times in PubMed abstracts.

Although TP53 was the most mentioned gene associated with CRC described in PubMed abstracts, this gene was not mentioned on the cancer.gov website, but it was found on expert curated sources. APC genes (APC, APC2, APCDD1) and MLH1 gene were found on all 3 resources (PubMed, curated sources and the cancer.gov website). KRAS, TGFBR2, and PPARG genes were identified on PubMed and curated sources but not on cancer.gov.

Genes described on the cancer.gov website that are known in theory as risk factors for CRC have been found on PubMed too. MUTYH was described in association with CRC in 7 abstracts, MSH2 was not described in the same abstract with CRC, but it was found in abstracts in association with other conditions that could be considered risk factors for CRC (ulcerative colitis, etc.). MSH6 was identified in 25 abstracts associated with CRC, PMS2 in 11, EPCAM in 10, POLD1 in 4, BMPR1A in 3, SMAD4 in 15, PTEN in 25 and STK11 in 8 PubMed abstracts. POLE and NTHL1 were not found in association with CRC in PubMed.

In table 1, we present APC genes associated with comorbidities of CRC and the number of abstracts in PubMed where these genes were described in relation to comorbidities of CRC.

Table 1. APC genes associated with CRC and comorbidities.

No	Diseases	ICD 9	PubMed abstracts
1	AD	331	6
2	Amyloidosis	277.3	15
3	Rheumatoid arthritis	714	3
4	Barrett's esophagus	530.85	8
5	Malignant neoplasm of bladder	188	5
6	Brain neoplasm	239.6	4
7	Burkitt's tumor or lymphoma	200.2	3
8	Celiac disease	579	2
9	Influenza	487	2
10	Leukemia	208	2

11	Megakaryocytic leukemia	207.2	3
12	Systemic lupus erythematosus	710	4
13	Malignant neoplasm of stomach	151	20
14	Multiple sclerosis	140	2
15	Nasal polyps	471	2
16	Obesity	278	2
17	Atrial premature beats	427.61	53
18	Septicemia	38.9	3
19	Hypogammaglobulinemia	279	3
20	Ulcerative colitis	554	4
21	Malignant neoplasm of mouth, unspecified	145.9	2
22	Malignant neoplasm of duodenum	152	2
23	Malignant neoplasm of gallbladder	156	2
24	Diabetes mellitus	250	3
25	Endometrial hyperplasia	621.3	2
26	Sebaceous cyst	706.2	3
27	Neuroendocrine tumors	209	2
28	Sepsis	995.91	2
29	Malignant neoplasm of intestinal tract, part unspecified	159	4
30	Graft-versus-host disease	279.5	3
31	Malignant neoplasm of pancreas	157	8
32	Malignant neoplasm of prostate	185	18
33	Malignant neoplasm of esophagus	150	5
34	Secondary and unspecified malignant neoplasm of lymph nodes	196	3
35	Malignant neoplasm of ovary	183	6
36	Candidiasis of mouth	112	3
37	Lymphoid leukemia, chronic	204.1	2

Presented are numbers of abstracts in PubMed where these associations were described. Comorbidities that are associated with CRC and APC in 2 or more abstracts are given in the table.

Associations between CRC and comorbidities with gene TP53 are shown in table 2. Only comorbidities that are in relationships with TP53 genes in 12 or more abstracts are given in the table.

Table 2. TP53 genes associated with CRC and comorbidities.

No	Disease	ICD 9	PubMed abstracts
1	Malignant neoplasm of esophagus	150	76
2	Nodular lymphoma	202	18
3	Leukemia of unspecified cell type, acute	208	22
4	Alzheimer's disease	331	20
5	Rheumatoid arthritis	714	29
6	Barrett's esophagus	530.85	38
7	Malignant neoplasm of bladder	180	159

8	Brain neoplasm NOS	239.6	87
9	Burkitt's tumor or lymphoma	200.2	53
10	Malignant neoplasm of thyroid gland	193	27
11	Malignant neoplasm of cervix uteri	180	130
12	ACUTE RHEUMATIC FEVER	390-392	16
13	Ulcerative colitis	556	36
14	Malignant melanoma of skin	172	19
15	Malignant neoplasm of liver	155.2	27
16	Malignant neoplasm of mouth, unspecified	145.9	51
17	Malignant neoplasm of stomach	151.9	167
18	Malignant neoplasm of pancreas	157	134
19	Endometriosis	617	19
20	Kaposi's sarcoma	176	12
21	Heart failure + Congestive heart failure	428	17
22	Malignant neoplasm of testis	186	12
23	Unspecified viral hepatitis C	70.7	23
24	Hodgkin's disease	201	50
25	Anaplastic large cell lymphoma	200.6	14
26	Unspecified immunity deficiency	279.3	13
27	Actinic keratosis	702	25
28	Acute kidney failure	584	16
29	Multiple myeloma	203	68
30	Unspecified disease of larynx	478.7	14
31	Leukemia of unspecified cell type	208	112
32	Lymphoid leukemia, chronic	204.1	248
33	Acute respiratory failure	518.81	16
34	Lymphoid leukemia, acute	204	51
35	Myeloid leukemia, acute	205	192
36	Myeloid leukemia, chronic	205.1	62
37	Mantle cell lymphoma	200.4	29
38	Malignant neoplasm of prostate	185	263
39	Neurofibromatosis, type 1 [von Recklinghausen's disease]	237.71	21
40	Malignant neoplasm of liver, secondary	197.7	31
41	Secondary and unspecified malignant neoplasm of lymph nodes	196	82
42	Malignant neoplasm of ovary	183	239
43	Myelodysplastic syndrome, unspecified	238.75	76

Presented are numbers of abstracts in PubMed where these associations were described. Comorbidities that are associated with CRC and TP53 in 12 or more abstracts are given in the table.

In table 3, associations between CRC, genes KRAS and comorbidities are presented.

Comorbidities that are described in relationships with these genes 2 or more times are shown.

Table 3. KRAS genes associated with CRC and comorbidities.

No	Disease	ICD 9	PubMed abstracts
1	Myeloid leukemia, acute	205	10
2	Malignant neoplasm of bladder	188	2
3	Malignant neoplasm of stomach	151	26
4	Malignant neoplasm of thyroid gland	193	5
5	Multiple myeloma	203	5
6	Obesity	278	2
7	Diseases of pancreas	577	3
8	Ulcerative colitis	556	7
9	Dengue	61	6
10	Pleural effusion	511	4
11	Patent ductus arteriosus	747	7
12	Atrial premature beats	427.61	5
13	Endometriosis	617	5
14	Uterine leiomyoma	218	4
15	Rash and other nonspecific skin eruption	782.1	3
16	Chronic pancreatitis	577.1	7
17	Unspecified viral hepatitis C	70.7	25
18	Leukemia of unspecified cell type	208	5
19	Lymphoid leukemia, acute	204	4
20	Malignant neoplasm of mouth, unspecified	145.9	2
21	Malignant neoplasm of gallbladder	156	3
22	Malignant neoplasm of uterus, part unspecified	179	2
23	Secondary malignant neoplasm of lung	197	14
24	Secondary malignant neoplasm of bone and bone marrow	198.5	3
25	Neurofibromatosis	237.7	2
26	Malignant neoplasm of pancreas	157	123
27	Malignant neoplasm of prostate	185	8
28	Malignant neoplasm of liver, secondary	197.7	27
29	Malignant neoplasm of esophagus	150	2
30	Secondary and unspecified malignant neoplasm of lymph nodes	196	14
31	Malignant neoplasm of biliary tract, part unspecified site	156.9	5
32	Malignant neoplasm of ovary	183	17

Presented are numbers of abstracts in PubMed where these associations were described. Comorbidities that are associated with CRC and KRAS in 2 or more abstracts are given in the table.

In table 4, association between CRC, MLH1 genes and comorbidities that are described in 2 or more PubMed abstracts are given.

Table 4. MLH1 genes associated with CRC and comorbidities.

No	Disease	ICD 9	PubMed abstracts
1	Azoospermia	606	2
2	Barrett's esophagus	530.85	2
3	Malignant neoplasm of bladder	188	7
4	Brain neoplasm	239.6	2
5	Ulcerative colitis	556	6
6	Endometrial hyperplasia	621.3	2
7	Endometriosis	617	2
8	Huntington's chorea	333.4	3
9	Leukemia of unspecified cell type	208	10
10	Lymphoid leukemia, chronic	204.1	2
11	Lymphoid leukemia, acute	204	2
12	Myeloid leukemia, acute	205	2
13	Nodular lymphoma	202	3
14	Malignant neoplasm of stomach	151	38
15	von Recklinghausen's disease	237.71	5
16	Atrial premature beats	427.61	2
17	Trematode infection NOS	121.9	2
18	Malignant melanoma of skin	172	2
19	Malignant neoplasm of mouth, unspecified	145.9	2
20	Malignant neoplasm of uterus, part unspecified	179	3
21	Malignant neoplasm of pancreas	157	3
22	Endometrial hyperplasia with atypia	621.33	3
23	Malignant neoplasm of prostate	185	7
24	Secondary and unspecified malignant neoplasm of lymph nodes	196	5
25	Malignant neoplasm of ovary	183	24

Presented are numbers of abstracts in PubMed where these associations were described. Comorbidities that are associated with CRC and MLH1 in 2 or more abstracts are given in the table.

In table 5 we present association between CRC, TGFBR2 genes and comorbidities described in 2 or more abstracts in PubMed.

Table 5. TGFBR2 genes associated with CRC and comorbidities.

No	Disease	ICD 9	PubMed abstracts
1	Alzheimer's disease	331	2
2	Aneurysm	442.9	4
3	Asthma	493	2
4	Ulcerative colitis	556	2
5	Patent ductus arteriosus	747	2
6	Ehlers-Danlos syndrome	756.83	2
7	Leprosy	30.9	2
8	Malignant neoplasm of stomach	151	10

9	Marfan syndrome	759.82	30
10	Systemic sclerosis	710.1	2
11	Migraine	346.9	2
12	Malignant neoplasm of mouth, unspecified	145.9	2
13	Malignant neoplasm of gallbladder	156	2
14	Dissection of aorta, unspecified site	441	2
15	Malignant neoplasm of pancreas	157	7
16	Malignant neoplasm of prostate	185	8
17	Secondary and unspecified malignant neoplasm of lymph nodes	196	2

Presented are numbers of abstracts in PubMed where these associations were described. Comorbidities that are associated with CRC and TGFBR2 in 2 or more abstracts are given in the table.

In table 6 we show association between CRC, PPARG genes and comorbidities that are described in 4 or more abstracts in PubMed.

Table 6. PPARG genes associated with CRC and comorbidities.

No	Disease	ICD 9	PubMed abstracts
1	Alzheimer's disease	331	34
2	Amyloidosis	277.3	6
3	Amyotrophic lateral sclerosis	335.2	10
4	Asthma	493	4
5	Atherosclerosis	440	37
6	Malignant neoplasm of bladder	188	7
7	Malignant neoplasm of thyroid gland	193	13
8	Cardiovascular Diseases	429.2	12
9	Coronary atherosclerosis	414	14
10	Cystic fibrosis	277	4
11	Malignant neoplasm of cervix uteri	180	5
12	Ulcerative colitis	556	7
13	Diabetes mellitus	250	100
14	Diabetes with renal manifestations	250.4	11
15	Diabetic retinopathy	362	12
16	Endometriosis	617	4
17	Heart failure	428	27
18	Huntington's chorea	333.4	21
19	HYPERTENSIVE DISEASE	401-405	31
20	End stage renal disease	585.6	5
21	Myeloid leukemia, chronic	205.1	5
22	Lipodystrophy	272.6	14
23	Malignant neoplasm of stomach	151	12
24	Obesity	278	208
25	Osteoarthritis, unspecified whether generalized or localized	715.9	11
26	Parkinson's disease	332	11
27	Polycystic ovaries	256.4	17
28	Retinal disorder	362.9	4

29	Schizophrenic disorders	295	7
30	Essential hypertension	401	5
31	Sepsis	995.91	4
32	Malignant neoplasm of liver, not specified as primary or secondary	155.2	5
33	Malignant neoplasm of pancreas	157	10
34	Malignant neoplasm of prostate	185	33
35	Dysmetabolic syndrome x	277.7	35
36	Cardiomyopathy	425	5

Presented are numbers of abstracts in PubMed where these associations were described. Comorbidities that are associated with CRC and PPARG in 4 or more abstracts are given in the table.

In table 7, given are 96 genes that are associated with CRC and present in both PubMed and expert curated sources.

Table 7. Ninety six genes associated with CRC and present in both: PubMed and curated sources.

ALDH1B1	CHEK2	IGF2	MTHFR	PTP4A3	SULT1A1
ALOX5	CTNNB1	IL1B	MUTYH	PTPN12	TAGLN
APC	CXCL10	IL32	MYC	PTPRJ	TCF7L2
ASCL2	DMBT1	JUN	NDRG2	RAD54B	TGFBR2
ATP7A	DNMT1	KCNH2	NFKB1	RCOR1	TNF
AXIN2	DPYD	KDM1A	NOS2	RECK	TNFRSF10A
BAX	EGFR	KRAS	NOTCH1	S100A4	TNFSF10
BCL2	ERBB2	LEF1	NOX1	SERPINB5	TP53
BECN1	FBP1	LEP	NOX4	SERTAD1	TP63
BIRC5	GAST	LGR5	ODC1	SLC2A1	TPM3
BRAF	GCG	MCM2	PCNA	SLC5A8	TYMP
BRD4	HES1	MECOM	POLB	SLCO1B3	TYMS
CBR1	HMGCS2	MIR98	PPARG	SOD2	VEGFA
CCAT1	HSPB1	MLH1	PRKN	SPARC	WT1
CCND1	ICAM1	MMP7	PROM1	SRC	XAF1
CDKN1A	IFNG	MMP9	PTGS2	STAT3	YBX1

Discussion

There is a compelling need to analyze genetics, comorbidities and procedures associated with CRC as very important risk and prognostic factors in the development and evolution of the cancer disease.

Findings of our research show that most of the top 30 procedures on ranked lists in early stages were the same in both genders. The most common procedure in both genders was transfusion of packed cells, followed by few radical surgical colorectal procedures. Most of the procedures in advanced stages were related to the treatment of the cancer diseases. The most common procedures among top 30 in advanced stages are almost identical between genders.

Analysis of procedures performed in patients with CRC discovers that certain cheaper diagnostic procedures (ultrasound) could be utilized more often and maybe replace more expensive diagnostic procedures (CT) in appropriate situations, which would decrease the cost of already expensive diagnostics and treatment of the cancer. The study also discovered that certain less invasive procedures (laparoscopic) could be used more often than open invasive procedures, which would lead to shorter hospital stays, faster recovery and potentially less infections and postoperative adhesions.

Venn diagrams and ranked lists show that around 80% of the top 30 comorbidities are identical in early and advanced stages of CRC in both genders, with very similar probabilities of occurrence in early and advanced stages. We calculated centrality measures with the goal to make inferences about significance of individual comorbidities. Considering findings of network analyses, RR was a more informative metric than Pearson's correlations. Evaluation of centrality measures used in this research discovered that both betweenness and node degrees come as valuable yet different measures considering findings of the most significant comorbidities associated with CRC.

Results of the comorbidity study were used to carry out analysis of associations between genes, CRC and comorbidities. We performed a text mining search of PubMed, using BeFree to identify genes associated with CRC, and comorbidities of CRC. We compared our findings to expert curated sources (DiSgeNet) which has known associations between genes and diseases, as well as theoretically known associations described on the cancer.gov website. Few genes (MSH2, POLE and NTHL1) associated with CRC, described on the cancer.gov website, as well as 54 genes associated with CRC in expert curated sources, were not found in PubMed abstracts. Since, we text mined only PubMed abstracts, it's possible that these associations were described in a full text of articles. We extracted 1,940 different genes from PubMed, associated with CRC. We found 150 different genes associated with CRC in curated sources. There were 1,790 genes more on PubMed than on DisGeNet sources, which is a novel contribution of our research. Even if we disregard genes – CRC associations described in only one abstract (1,160 – 60%) and consider associations described in 2 or more abstracts, there are still 630 more genes-CRC associations discovered on PubMed than on curated sources. Ninety six genes were identical on both sources, meaning they were verified by experts as genes associated with CRC. The rest of discovered genes in associations with CRC need further study by experts, in order to confirm their relevance. We included only genes-comorbidities associations described two or more times in PubMed abstracts, which potentially suggest real associations that should be further investigated and entered into expert curated sources.

Genetic findings could be used in identification of more individuals who would benefit from genetics consultations. This would lead to improvement of guidelines for genetic testing and discovery of patients that are at increased risk to develop CRC in the future.

Conclusion

Numerous studies point out that the presence of comorbidities weighs as a negative factor that contributes to faster deterioration of one's health. Our findings suggest which comorbidities should be highly expected along the course of the cancer disease. Preventive measures and early diagnostics of these comorbidities could be applied in order to prevent comorbidities or discover them early. Prevention, early discovery, and treatment of comorbidities would potentially lead to slower progression and better outcomes of CRC.

In our research, we attempted to elevate the significance of big data accumulated in PubMed, which is a rich resource of medical scientific publications, to extract information about associations between genes and diseases. This type of data mining of PubMed and other large databases will require expert knowledge to analyze and sort results and evaluate correctness of extracted associations between medical concepts in order to store them into scientifically verified databases. Results of the study could be considered as a novelty contributing to better understanding of risk factors (genes, comorbidities) to the onset and development of CRC. Genes associated with CRC and comorbidities, found only in PubMed abstracts should be evaluated as a potential risk or predictive factor for development of these medical conditions. Subsequent research is necessary to evaluate genes-CRC-comorbidities associations and incorporate findings into expert domain-specific curated databases.

Abbreviations

CRC: Colorectal cancer

HCUP: Healthcare Cost and Utilization Project

586 **SID:** State Inpatient Databases
587 **DisGeNet:** database integrates information of human disease-gene associations
588 **BeFree:** Bio Entity Finder and RElation Extraction
589 **RR:** Relative risk
590 **UTI:** Urinary tract infection
591 **CHF:** Congestive heart failure
592 **CAD:** Coronary artery disease
593 **CKD:** Chronic kidney disease
594 **AKF:** Acute kidney failure

595

596

597 **Declarations**

598 **Acknowledgements**

599 The authors gratefully acknowledge the support of the King Abdullah University of Science and
600 Technology's Center Partnership Fund Program. Healthcare Cost and Utilization Project (HCUP),
601 Agency for Healthcare Research and Quality provided data used in this study. The authors also
602 thank Stefan Obradovic for proofreading and editing the language of the manuscript.

603

604 **Authors' contributions**

605 BL, MP, ZO, SR, JA, VB, CVN designed the project. BL wrote the manuscript. BL, MP, JA
606 searched and downloaded data from HCUP. CVN, VB text mined PubMed and DisGeNet. BL
607 performed comorbidity network analysis. BL, SR, VB and CVN completed genes-diseases

608 associations analysis. ZO performed writing-review and editing. All authors read and approved
609 the final manuscript.

610 **Ethics approval and consent to participate**

611 Not applicable

612 **Consent for publication**

613 Not applicable

614 **Competing interests**

615 The authors declare that there are no competing interests regarding the publication of this paper

616 **Availability of data and material**

617 The datasets generated and analyzed in the comorbidity part of this study are available in the HCUP

618 SID database: <https://www.hcup-us.ahrq.gov/db/state/siddbdocumentation.jsp>.

619 Data analyzed in the genes part of the study were extracted by the text mining from the PubMed:

620 <https://www.ncbi.nlm.nih.gov/pubmed/>.

621

622 **Funding**

623 This research is partly sponsored by the King Abdullah University of Science and Technology's
624 Center Partnership Fund Program.

625

626

627 **References**

628 1. Kumar V, Abbas A, Fausto N, et al. Robbins and Cotran Pathologic basis of diseases. 9th edition.
629 2014.

630 2. Survival Rates for Colorectal Cancer, by Stage. [https://www.cancer.org/cancer/colon-rectal-](https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/survival-rates.html)
631 [cancer/detection-diagnosis-staging/survival-rates.html](https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/survival-rates.html)

632 3. Allen-Mersh TG, Earlam S, Fordy C, et al. Quality of life and survival with continuous hepatic-
633 artery floxuridine infusion for colorectal liver metastases. *Lancet* 1994;344(8932):1255-60.

634 4. Oines M, Helsingen LM, Bretthauer M, et al. Epidemiology and risk factors of colorectal polyps.
635 *Best Pract Res Clin Gastroenterol* 2017;31(4):419-24.

636 5. Goldblum JR, Lamps MD LW, McKenney JK, et al. *Rosai and Ackerman's Surgical Pathology*.
637 Elsevier 2018:648-843.

638 6. Hidalgo CA, Blumm N, Barabasi AL, et al. A Dynamic Network Approach for the Study of
639 Human Phenotypes. *PLoS Computational Biology* 2009;5(4):e1000353.

640 7. Rual JF, Venkatesan K, Hao T, et al. Towards a proteome-scale map of the human protein-
641 protein interaction network. *Nature* 2005;7062:1173-8.

642 8. Grundner J, Prokosch HU, Sturzl M, et al. Predicting Clinical Outcomes in Colorectal Cancer
643 Using Machine Learning. *Stud Health Technol Inform* 2018;247:101-5.

644 9. Bychkov D, Linder N, Turkki R, et al. Deep learning based tissue analysis predicts outcome in
645 colorectal cancer. *Scientific reports*. 2018;8:3395 | DOI:10.1038/s41598-018-21758-3

646 10. Schneider NI, Langner C. Prognostic stratification of colorectal cancer patients: Current
647 perspectives. *Cancer Manag Res* 2014;6:291–300.

648 11. Genetics of colorectal cancer. [https://www.cancer.gov/types/colorectal/hp/colorectal-](https://www.cancer.gov/types/colorectal/hp/colorectal-genetics-pdq#link/_89)
649 [genetics-pdq#link/_89](https://www.cancer.gov/types/colorectal/hp/colorectal-genetics-pdq#link/_89).

650 12. Kanth P, Grimmett J, Champine M, et al.: Hereditary Colorectal Polyposis and Cancer
651 Syndromes: A Primer on Diagnosis and Management. *Am J Gastroenterol* 2017;112 (10): 1509-
652 1525.

653 13. Bravo A, Piñero J, Queralt-Rosinach N, et al. Extraction of relations between genes and
654 diseases from text and large-scale data analysis: implications for translational research. *BMC*
655 *Bioinformatics* (2015) 16:55. DOI 10.1186/s12859-015-0472-9.

656 14. Pinero J, Bravo A, Queralt-Rosinach N, et al. DisGeNET: a comprehensive platform
657 integrating information on human disease-associated genes and variants. *Nucleic Acids*
658 *Research*, 2017, Vol. 45, Database issue D833–D839. doi: 10.1093/nar/gkw943.

659 15. Williams GR, Mackenzie A, Magnuson A, et al. Comorbidity in Older Adults with Cancer. *J*
660 *Geriatr Oncol* 2016;7(4): 249–57.

661 16. Boyd CM, Darer J, Boult C, et al. Clinical practice guidelines and quality of care for older
662 patients with multiple comorbid diseases: implications for pay for performance. *JAMA* 2005;
663 294(6):716–24.

664 17. Sogaard M, Thomsen RW, Bossen KS, et al. The impact of comorbidity on cancer survival: a
665 review. *Clinical epidemiology*. 2013; 5(Suppl 1):3–29.

666 18. van Eeghen EE, Bakker SD, van Bochove A, et al. Impact of age and comorbidity on survival
667 in colorectal cancer. *J Gastrointest Oncol* 2015;6(6):605-12

668 19. Ostensfeld EB, Nørgaard M, Thomsen RW, et al. Comorbidity and survival of Danish patients
669 with colon and rectal cancer from 2000-2011: a population-based cohort study. *Clin Epidemiol*
670 2013;5(Suppl 1):65-74.

671 20. Rustgi AK. The genetics of hereditary colon cancer. *Genes Dev* 2007;21(20): 2525-38.

672 21. Barnetson RA, Tenesa A, Farrington SM, et al.: Identification and survival of carriers of
673 mutations in DNA mismatch-repair genes in colon cancer. *N Engl J Med* 2006;354 (26):2751-63.

- 674 22. Hampel H, Bennett RL, Buchanan A, et al. A practice guideline from the American College of
675 Medical Genetics and Genomics and the National Society of Genetic Counselors: referral
676 indications for cancer predisposition assessment. *Genet Med* 2015;17 (1): 70-87.
- 677 23. I Munteanu I, Mastalier B. Genetics of colorectal cancer. *J Med Life*, 2014;7(4):507–11.
- 678 24. Obuch JC, Ahnen DJ. Colorectal Cancer: Genetics is Changing Everything. *Gastroenterol*
679 *Clin North Am*, 2016;45(3):459-76. doi: 10.1016/j.gtc.2016.04.005.
- 680 25. Lee KC, Chung KC, Chen HH, et al. Short-term postoperative outcomes of colorectal cancer
681 among patients with chronic liver disease: a national population-based study. *BMJ Open*.
682 2018;8(7):e020511. doi:10.1136/bmjopen-2017-020511.
- 683 26. Ljubic B, Gligorijevic Dj, Gligorijevic J, et al. Social network analysis for better understanding
684 of influenza. *J Biomed Inform*, 2019;93:103161. <https://doi.org/10.1016/j.jbi.2019.103161>.

Figures

Figure 1

Comparative Venn diagrams of the top 30 procedures in females and males.

Figure 2

Comparative Venn diagrams of comorbidities in females and males.

Figure 3

Networks of comorbidities in early stages of colorectal cancer in females

Figure 4

Networks of comorbidities in advanced stages of colorectal cancer in females

Figure 5

Networks of comorbidities in early stages of colorectal cancer in males.

Figure 6

Networks of comorbidities in advanced stages of colorectal cancer in males.