SCIENTIFIC REPORTS
natureresearch

Check for updates

OPEN

PCOSKB$_{R2}$: a database of genes, diseases, pathways, and networks associated with polycystic ovary syndrome

Mridula Sharma[1,2], Ram Shankar Barai[1,2], Indra Kundu[1,2], Sameeksha Bhaye[1], Khushal Pokar[1] & Susan Idicula-Thomas[1]✉

PolyCystic Ovary Syndrome KnowledgeBase (PCOSKB$_{R2}$) is a manually curated database with information on 533 genes, 145 SNPs, 29 miRNAs, 1,150 pathways, and 1,237 diseases associated with PCOS. This data has been retrieved based on evidence gleaned by critically reviewing literature and related records available for PCOS in databases such as KEGG, DisGeNET, OMIM, GO, Reactome, STRING, and dbSNP. Since PCOS is associated with multiple genes and comorbidities, data mining algorithms for comorbidity prediction and identification of enriched pathways and hub genes are integrated in PCOSKB$_{R2}$, making it an ideal research platform for PCOS. PCOSKB$_{R2}$ is freely accessible at http://www.pcoskb.bicnirrh.res.in/.

Polycystic ovary syndrome (PCOS) is the most common endocrine disorder in women of reproductive age[1]. The syndrome encompasses a broad spectrum of signs and symptoms, making the diagnosis of PCOS challenging. There exist many society-based guidelines for PCOS diagnosis such as the (i) Rotterdam criteria accepted by European Society for Human Reproduction and Embryology(ESHRE) and American Society for Reproductive Medicine (ASRM)[2]; (ii) National Institutes of Health or National Institute of Child Health and Human Disease (NIH/NICHD) criteria[3] and (iii) Androgen Excess and PCOS Society (AE-PCOS/AES) criteria[4]. These guidelines rely on the presence of oligo-anovulation and hyperandrogenism, after excluding other androgen excess or related disorders, for diagnosis of PCOS. The prevalence of PCOS globally ranges from 2.2 to 26% contingent upon the population assessed and the criteria used for evaluation[5]. Many of the women with PCOS suffer from various comorbid conditions such as glucose intolerance[6], type-II diabetes[7], cardiovascular ailments[8], anxiety disorders[9], bipolar disorders[10] and sleep-related disorders[11].

The increasing prevalence of PCOS and its profound impact on the physical and mental health of women has catapulted research efforts to elucidate the genetic etiology and pathophysiology of PCOS[12]. This, in turn, has led to a surge in PCOS-related data available in the public domain; calling for an urgent need to manually curate and collate this information as online databases for researchers and clinicians.

The databases dedicated to PCOS, currently available online are PCOSKB[13] and PCOSBase[14]. As on date, PCOSDB[15] is not accessible. PCOSBase, categorized as a manually curated database, lists 8,185 proteins as associated with PCOS. This data is a compilation from 9 databases and 30 published expression studies, without having stringent criteria for cataloguing a protein as "PCOS-related". PCOSKB, developed by our group in 2015, was created by critically reviewing the scientific literature available for PCOS. The manual curation exercise resulted in a list of 241 genes, which was further linked with relevant molecular, biochemical, and clinical data along with supporting reference literature.

Over the past 5 years, there has been a significant increase in the data available on PCOS. Here, we present an update to the content and functionality of the PCOSKB database. PCOSKB$_{R2}$ holds information of 533 genes and 29 miRNAs (manually curated) identified from peer-reviewed literature, based on experiments such as RT-PCR, western blotting, immunochemistry, and cell-based assays. Additionally, information on 4,023 genes identified

[1]Biomedical Informatics Center, Indian Council of Medical Research-National Institute for Research in Reproductive Health, Mumbai 400012, India. [2]These authors contributed equally: Mridula Sharma, Ram Shankar Barai and Indra Kundu. ✉email: thomass@nirrh.res.in

from microarray expression studies on PCOS is also included in PCOSKB$_{R2}$. The PCOS genes are further linked with gene ontology terms, pathways, diseases, and SNPs.

Besides retrieving data, researchers can analyse the data in PCOSKB$_{R2,}$ using various tools embedded in the database such as *Comorbidity analysis* for estimating the risk of diseases to co-occur with PCOS; *Network analysis* for identifying enriched pathways and hub genes and *Venn analysis*[16] for finding common and unique genes, pathways and ontologies. PCOSKB$_{R2}$ will enable researchers and clinicians to efficiently interrogate the published data on PCOS and identify gaps in our current understanding of PCOS and its comorbidities.

Results and discussion

PCOSKB$_{R2}$ was developed using PHP 7.2.24, MariaDB Server 10.1.44, JavaScript, AnyChart 8.7.1, vis.js 4.21, R version 3.6.3 and XHTML 1.0. It has client server-based architecture and is hosted on Apache webserver 2.4.29 with a Linux environment.

PCOSKB$_{R2}$ has an interactive and user-friendly interface. The homepage provides a short description of the database and its functionalities. The data is organized into datasets dedicated to (a) genes, (b) miRNAs, (c) SNPs, (d) diseases, (e) pathways, and (f) gene ontology terms associated with PCOS (Fig. 1a,b). These datasets can be easily accessed using the navigation tabs located on the top panel of the webpage. A brief description of these tabs is given below:

- *Search*

  a. *Quick search* enables users to retrieve information based on keywords; all the information available in PCOSKB related to the keyword is displayed.
  b. *Advanced search* enables users to build specific queries for a gene, protein, SNP, miRNA, diseases, or pathways associated with PCOS.

- *Browse* This tab enables users to surf the datasets for genes, miRNAs, SNPs, diseases, pathways, and gene ontology terms associated with PCOS.
- *Tools* Algorithms for comorbidity, network, and Venn analysis can be accessed here.

  a. *Comorbidity Analysis* This tool can be used to predict comorbidity for selected diseases based on (i) shared genes, (ii) uniqueness of shared genes, (iii) shared ontologies, and (iv) network-based separation of shared genes (Fig. 1c1). The results for each of these modules can be downloaded as heatmap images (colored based on comorbidity scores) and spreadsheets with detailed information on shared genes and pairwise comorbidity scores for the selected diseases.
  b. *Network analysis* The tool provides a disease-disease network for selected diseases, the enriched pathways in these diseases, and the hub and bottleneck genes that are critical for these diseases (Fig. 1c3). The results can be downloaded as spreadsheets or images.
  c. *Venn analysis* This tool can be used to illustrate the unique and/or common genes, pathways, and ontologies for 2 or more (up to 6) diseases (Fig. 1c2). The analysis can be downloaded as Venn images or spreadsheets.

- *Help*: This page provides detailed information, with examples, for efficiently navigating the PCOSKB interface and using the data-mining algorithms.

The applications of these datasets and algorithms for estimating the comorbidity risk and understanding the genetic and functional overlap in comorbid conditions of PCOS are demonstrated by case studies.

A. Estimation of comorbidity risk:

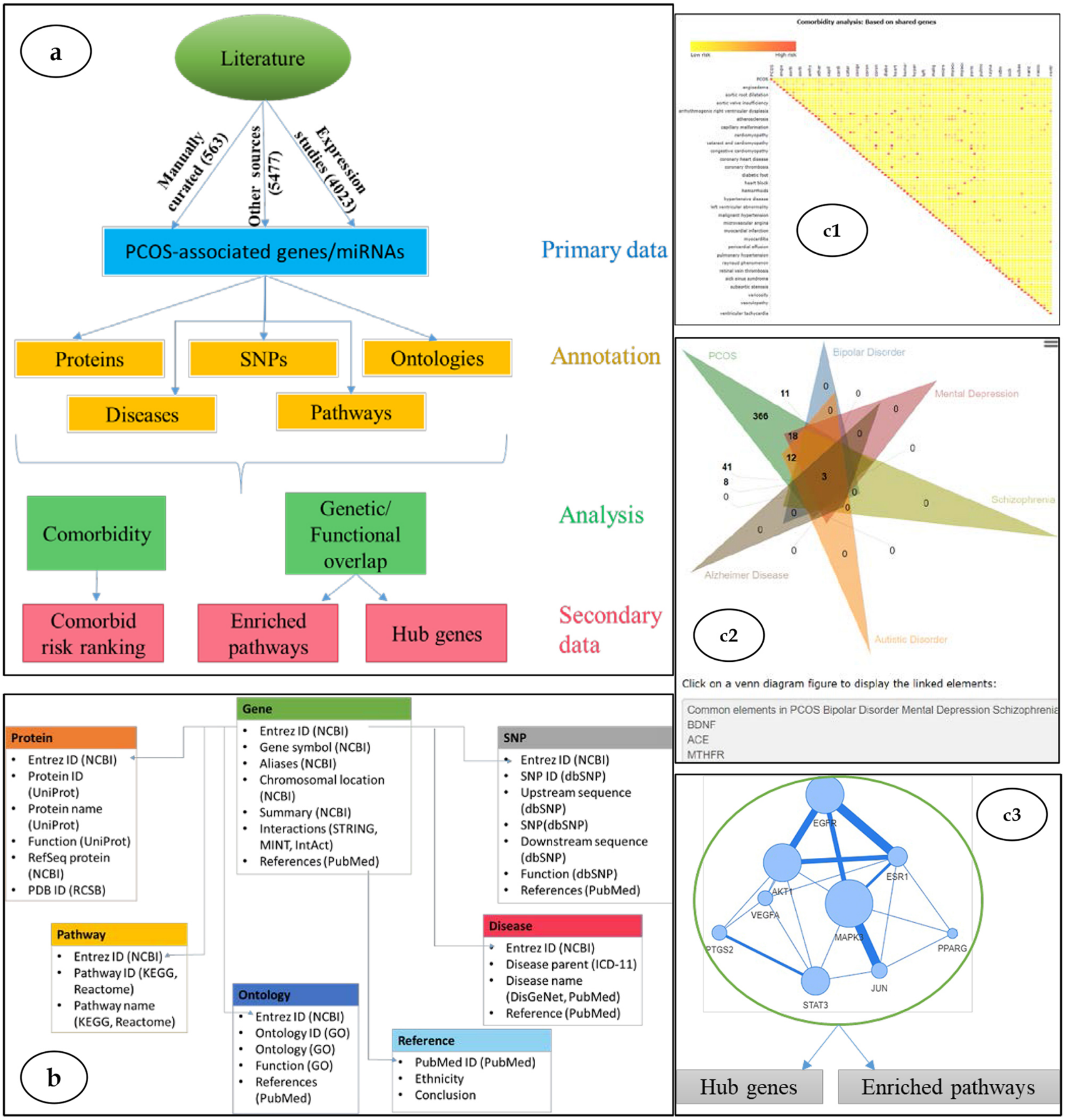
Case 1: *PCOS, Diabetes, and Hypertension.*

There is ample clinical evidence that women with PCOS are more likely to suffer from diabetes and hypertension as compared to other cardiac ailments[17–20].

The comorbidity risk can be estimated using the '*Comorbidity analysis*' algorithm in PCOSKB$_{R2}$. In accordance with the clinical reports, when disease terms such as diabetes mellitus, hypertensive diseases along with a less frequently observed comorbidity such as aortic diseases were analyzed for comorbidity scores; it was found that the risk for diabetes and hypertensive diseases to co-occur with PCOS was much higher as compared to aortic diseases. Expectedly, the maximum comorbidity score amongst the selected diseases was found to be between aortic diseases and hypertension (Fig. 2A). The above example illustrates the utility of the comorbidity analysis algorithm for estimating the risk of diseases to co-occur in PCOS.

Case 2: *PCOS and Psychological disorders.*

Women with PCOS are known to have an increased risk (albeit at varying levels) of suffering from mental health conditions such as anxiety, depression, and schizophrenia[21,22]. A study by Rassi et al., concluded that 57% of women with PCOS are diagnosed with at least one of the psychiatric disorders[23]. In an ambulatory population of 72 women with PCOS, it was observed that mental depression and schizophrenia were the most and least prevalent respectively among the psychiatric disorders[23]. Through a population-based retrospective study in a cohort of 5,431 women with PCOS and 21,724 controls, a significantly higher incidence of depressive and anxiety



Click on a venn diagram figure to display the linked elements:

Common elements in PCOS Bipolar Disorder Mental Depression Schizophrenia
BDNF
ACE
MTHFR

Figure 1. Conceptual and relational view of data and tools in PCOSKB$_{R2}$.

disorders were reported in women with PCOS[24]. In another study, the prevalence of psychiatric comorbidity and depression was reported as the most common disorder in women with PCOS followed by anxiety[25]. Meta-analysis of 57 studies (172,040 patients) summarised that women with PCOS were most likely to get diagnosed with depression followed by anxiety[26].

These clinical observations were accurately captured through the comorbidity scores generated using the network-based separation method. Mental depression had the highest comorbidity risk followed by anxiety disorders and schizophrenia (Fig. 2B). It is noteworthy that although maximum number of genes (124) overlapped between PCOS and schizophrenia, as reflected in the edge thickness between these 2 disease nodes; comorbidity analysis correctly estimated the least risk for comorbidity with schizophrenia amongst the three mental diseases, in accordance with literature reports; highlighting the predictive power of network-based separation method for comorbidity analysis.

B. Identification of the genetic and functional overlap in comorbid conditions.

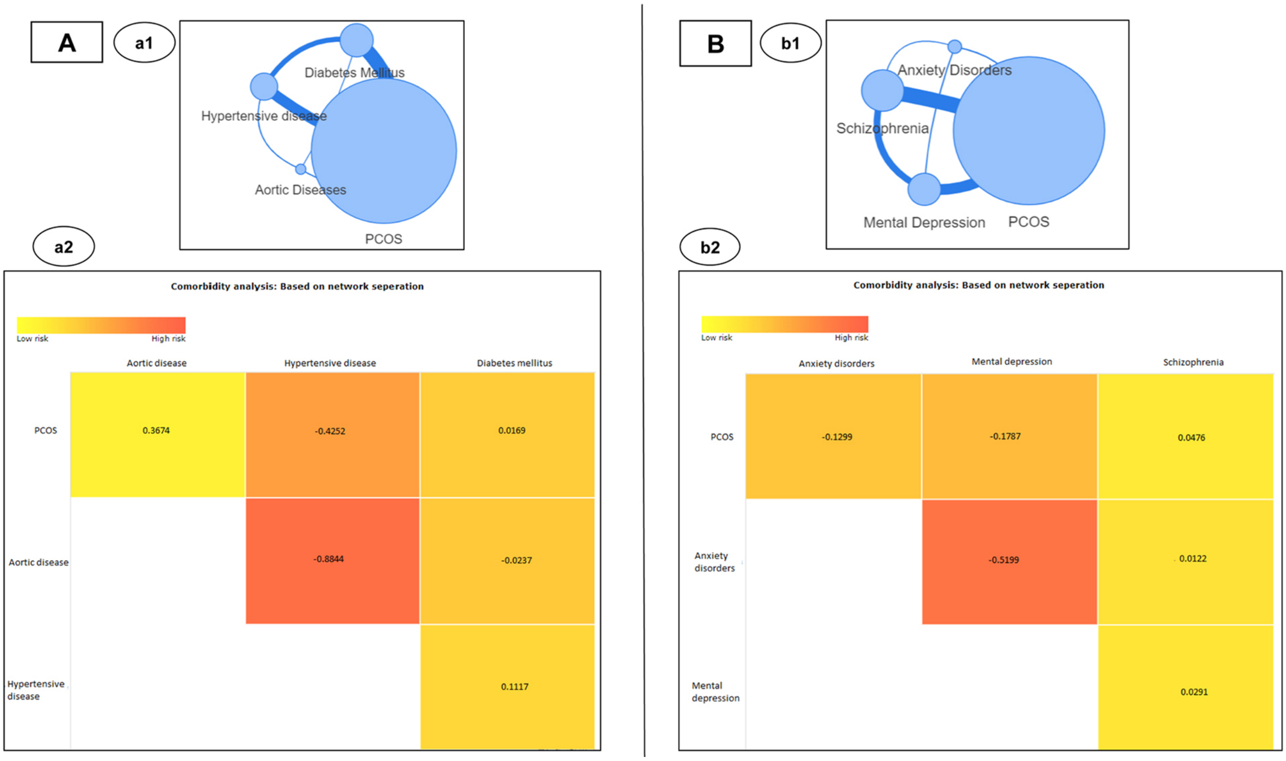


Figure 2. Network-based comorbidity analysis for PCOS and (**A**) diabetes and hypertension; (**B**) psychological disorders.

Case 1: *PCOS, Diabetes, and Hypertension.*

Although, diabetes and hypertension are commonly observed comorbid conditions in women with PCOS; not much is known about the genetic overlap of these disorders[27].

Venn analysis revealed that 32 genes and 364 pathways are commonly associated with PCOS, diabetes, and hypertension (Supplementary Table S1). Network analysis identified 104 enriched pathways, 21 hub genes, and 10 bottleneck genes for these diseases (Supplementary Figs. S1a1 and S1a2, Supplementary Table S1). Hub genes, due to their high degree of inter-cluster connectivity, play an important role in the crosstalk of enriched pathways. We mined literature for ascertaining the association of these 21 genes with the comorbid conditions of diabetes, hypertension, and PCOS. Of the 21 genes, we found literature evidence for association of four genes (*ESR1*, *PTGS2, LEP, PPARG)* with these comorbidities, as detailed below.

(i) *ESR1* codes for estrogen receptor alpha and hence *ESR1* mutations can increase the risk of estrogen-dependent pathophysiologies. In a study by Zhao L et al., *ESR1* polymorphisms were reported to be associated with hypertension and diabetes[28]. A case–control study by Jiao X et al., documented that altered expression of *ESR1* can influence the risk of PCOS and its upregulation may contribute to abnormal follicular development[29,30].

(ii) Prostaglandin-endoperoxide synthase (*PTGS2)* is a key enzyme for biosynthesis of the inflammatory hormone prostaglandin. It is known to be upregulated in granulosa cells of women with PCOS and arteries of patients with hypertension and diabetes[31,32].

(iii) Leptin hormone encoded by the leptin gene (*LEP)* plays an important role in the regulation of energy homeostasis and body weight management. Several independent studies have reported the association of leptin receptor deficiency in diabetes, hypertension, and PCOS. High circulatory leptin has been observed in patients with a cluster of metabolic syndrome including hypertension, diabetes[33], and PCOS[69].

(iv) Peroxisome proliferator-activated receptor gamma (*PPARG)* regulates adipocyte differentiation and thereby controls beta-oxidation of fatty acids. Mutations in *PPARG* are known to increase the risk for development of hypertension and diabetes[34].

In addition to the identification of hub and bottleneck genes, the *View interaction* option in the *Gene network analysis* tool can be used to display the tissue-specific interacting partners of each gene in the network (Supplementary Fig. S1). Using this feature, we identified two genes (*PON1, ADIPOQ)* that interact with multiple hub genes (Supplementary Figs. S1a3 and S1a4). *PON1* interacts with six hub genes (*TNF*, *IL6*, *INS*, *CCL2*, *LEP*, *PPARG*) and one bottleneck gene (*LIPC*) (Supplementary Fig. S1a4). Adiponectin *(ADIPOQ)* interacts with 19 hub genes that are expressed in adipose tissue (Supplementary Fig. S1a3). The association of both these genes in the comorbid conditions of type 2 diabetes, hypertension, and PCOS is documented in the literature. Paraoxonase-1 (*PON1*) mediates enzymatic protection of low-density lipoprotein (LDL) against oxidative modifications and is known to be associated with diabetes, hypertension, and PCOS[35,36]. Low levels of adiponectin are associated with several obesity-related disorders[37] and *ADIPOQ* is a biomarker for type-2 diabetes, hypertension[38], and PCOS[39].

This case study illustrates the utility of the *Gene network analysis* tool in deciphering the genetic and functional overlap of comorbid conditions. While the role of all the identified hub genes in PCOS, diabetes, and hypertension individually has been well established, it would be worthwhile to establish the role of these hub genes in the pathophysiology of PCOS, diabetes, and hypertension, as a combined disease state, and explore them as polypharmacological drug targets.

Case 2: *PCOS and Psychological disorders—anxiety and mental depression.*

Insulin resistance, obesity, and altered levels of androgens (Supplementary Table S2) have been reported as the common pathophysiological link between PCOS and psychiatric disorders[24,40]. Interestingly, evaluation of enriched pathways for the top two psychological disorders (mental depression and anxiety) that are comorbid with PCOS revealed pathways that represent these cellular mechanisms (Supplementary Table S2, Supplementary Figs. S1b1 and S1b2, Supplementary Table S1).

Network analysis of the enriched pathways revealed 21 hub genes and 10 bottleneck genes. Of these, the role of two hub genes (*IL6*, *STAT3*) in the comorbidity of PCOS and selected psychiatric disorders has been reported in literature. Kawamura S et al., reported elevated levels of inflammatory cytokine *IL6* in women suffering from PCOS and depression[41]. The negative association of *STAT3* with anxiety and depression have been reported by Feng and Shao in PCOS induced rat models[42]. Anxiety and depression in rats were analysed based on their decreased locomotor activity in behavioural tests such as open-field tests, object recognition tests, and elevated plus maze tests.

Case 3: *PCOS and Pregnancy-related disorders—preeclampsia.*

Women with PCOS are known to be at higher risk of pregnancy-related disorders as compared to women without PCOS[43,44]. In PCOSKB, genes, and miRNAs associated with pregnancy-related disease terms like "Pregnancy complications, Cardiovascular", "Pregnancy associated hypertension", "Ectopic pregnancy", "Gestational diabetes", and "Preeclampsia" can be accessed under the disease category of reproductive disorders.

miRNAs are known to play a critical role in the pathogenesis of PCOS and pregnancy-related disorders[45–47]. Pathways such as adipocytokine signaling, oxytocin signaling, TNF signaling, progesterone-mediated oocyte maturation, estrogen signaling, MAPK, and FoxO signaling are known to be regulated by miRNAs and associated with pregnancy outcome[48,49].

miRNA-based pathway enrichment analysis of preeclampsia revealed 88 enriched pathways that included progesterone-mediated oocyte maturation, estrogen signaling, MAPK signaling, and FoxO signaling pathways (Supplementary Table S1); these pathways are known to be associated with PCOS and preeclampsia in literature[49–51].

Conclusion and future directions

The aim of developing PCOSKB$_{R2}$ was to provide a one-stop online portal for accessing manually curated information on PCOS to the community of clinicians and researchers. The genes, listed in the manually curated dataset of PCOSKB$_{R2}$ were identified based on the inference and data mined from publications. Relevant annotations of these genes such as gene interactions, pathway associations, and SNPs have been provided along with links to the reference literature.

This second release of PCOSKB has substantial advancement both in terms of data and analysis tools[13]. In addition to the advanced search and browser features for efficiently interrogating the database, users can avail of the tools to predict comorbidity risks, enriched pathways, and hub genes for selected diseases. These tools are powerful for gaining insights on the comorbidities of PCOS and the underlying gene-pathway associations, as can be seen by the aforementioned case studies. However, users need to be aware and cautious of the publishing or literature bias that can lead to erroneous inferences.

The impact of publication bias on the results of the comorbidity analysis tool can be assessed by the following example. Women with PCOS are known to suffer from an increased risk of endometrial cancer followed by ovarian cancer as compared to women without PCOS[50]. The incidence of breast cancer is similar in women with and without PCOS[41,50,51]. The comorbidity analysis tool, using the method of shared genes, incorrectly predicted the highest risk of comorbidity for breast, followed by ovarian and least for endometrial cancer (Fig. 3). This error is inadvertently caused due to the positive publication bias for breast cancer (407,285 PubMed records) as compared to ovarian (116,514 PubMed records) and endometrial cancers (37,950 PubMed records). Hence, the genes that are known to be associated with endometrial cancer are far lesser (38 genes) than ovarian (57 genes) and breast cancers (129 genes).

The network separation based algorithm identified the highest comorbidity risk for ovarian, followed by breast and endometrial cancers (Fig. 3). The network separation method is based on the distance/separation of the disease-causing genes in pathway networks and therefore is more robust and less dependent (not independent) on the number of disease-causing genes as compared to the algorithm of shared genes. This algorithm should, therefore, be the choice for comorbidity prediction when a fewer number of diseases; with possibility for publication bias is analysed.

The incidence of PCOS is rising globally[52–56] and we expect the data, generated on PCOS, to increase exponentially in the years to come. Depending on the availability and nature of data generated from these research efforts, PCOSKB$_{R2}$ will be updated with new information and analysis tools. Hopefully, with more data, the negative impact of publication bias will be reduced. PCOSKB$_{R2}$ will be a comprehensive source of updated and curated information on gene-disease-pathway associations in PCOS and its comorbidities.

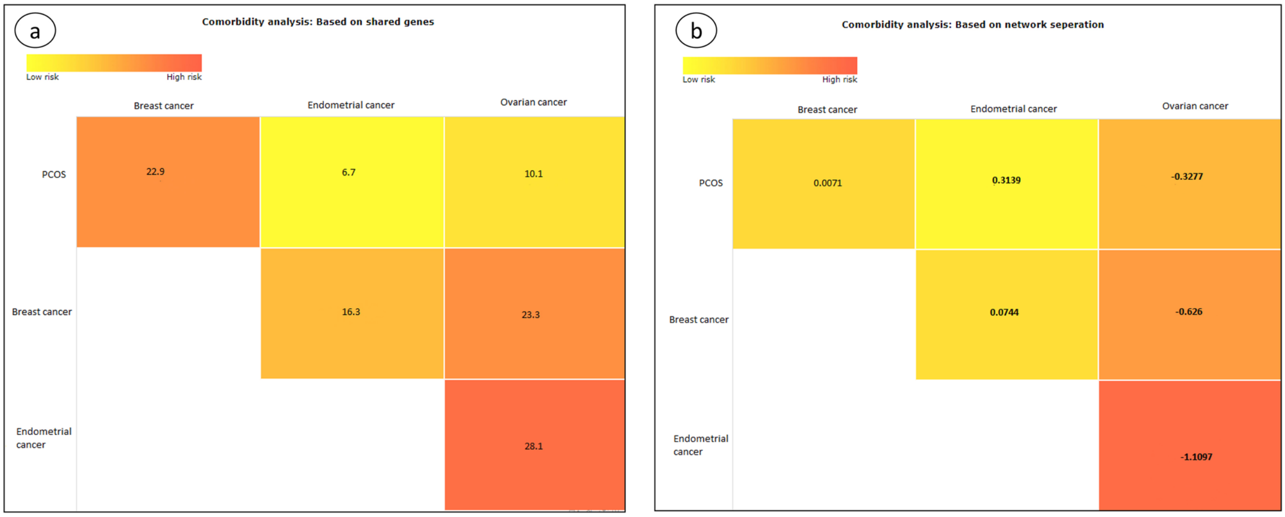


Figure 3. Comorbidity analysis for PCOS and cancers using (**a**) shared genes and (**b**) network-separation methods.

Methods

Dataset curation.

Curation of the gene dataset. The genes associated with PCOS were identified by querying PubMed[57] with MeSH(Medical Subject Headings)[58] terms such as, "Ovary Syndrome, Polycystic", "Syndrome, Polycystic Ovary", "Stein-Leventhal Syndrome", "Stein Leventhal Syndrome", "Syndrome, Stein-Leventhal", "Sclerocystic Ovarian Degeneration", "Ovarian Degeneration, Sclerocystic", "Sclerocystic Ovary Syndrome", "Polycystic Ovarian Syndrome", "Ovarian Syndrome, Polycystic", "Polycystic Ovary Syndrome 1", "Sclerocystic Ovaries", "Ovary, Sclerocystic", "Sclerocystic Ovary", "PCOS" and "Gene". Using this query, 1561 literature records were retrieved from PubMed.

The association of 533 genes with PCOS was manually confirmed by critically reviewing the 1561 publications. A gene was verified to be PCOS-associated if the literature mentions experimental evidence based on RT-PCR, western blotting, immunochemistry, and cell-based assays. Additional annotations such as nature of the study population, ethnicity, mutations/SNPs, unique identifiers for gene and protein records, protein structures, family and ontology details, metabolic pathway information were obtained from literature and mapping the gene records to databases such as Gene[59], dbSNP[60], Ensembl[61], UniProt[62], PDB[63], GO[64], KEGG[65], OMIM[66], Reactome[67] and STRING[68] (Supplementary Table S3).

Curation of the gene-disease association dataset. Disease associations of the PCOS genes were retrieved from DisGeNET[69] and PubMed[57] databases. The disease terms in DisGeNET that are linked to PubMed literature and have an active MedGen[70] *ConceptID (CUI)* were retained for further curation. The terms with disease type as *"phenotype"* and disease semantic type as "finding", "pathologic function", "sign or symptom", "injury or poisoning", "experimental model of disease", "experimental model of disease; Neoplastic process", "anatomical abnormality", "organism attribute" were discarded from the list as the terms under these headers did not refer to diseases.

This list was further subdivided into two sets based on the source of information in DisGeNET[69]. Dataset '*A*' comprised of gene-disease associations collated in DisGeNET from manually curated databases such as ClinVar[71], CTD[72], Genomics England[73], GWAS Catalog[74] and GWAS[75] and Dataset '*B*' had information collated from text mining datasets such as BEFREE[76] and LHGDN[77]. Since dataset '*A*' records were from curated sources, these were included in $\mathrm{PCOSKB_{R2}}$ without further verification. For dataset '*B*', gene-disease associations were validated based on rigorous manual curation. The associated literature was reviewed carefully and evidence for gene-disease association was sourced from experimental techniques involving human samples, such as RT-PCR, western blotting, immunochemistry, and cell-based assays. Genes that did not have any disease information in DisGeNET were queried in PubMed and publication records were mined using pubmed.mineR package[78].

In cases, wherein multiple disease terms referred to the same disease, the terms were retitled as explained in Table 1.

Unique categorization of disease groups. Many of the disease terms in DisGeNET[69] are mapped to multiple MeSH[58] headings. E.g. ovarian neoplasm is linked to neoplasms and reproductive disorders. An empirical rule-based method based on ICD-11[79] classification (Fig. 4) was adopted to uniquely categorize the disease terms at the parent level.

For complete documentation of merged terms refer to Supplementary Table S1.

Tools.

Comorbidity analysis. For a pair of diseases (D_i, D_j), the list of PCOS-associated genes was retrieved from the gene-disease dataset of $\mathrm{PCOSKB_{R2}}$ (see "Curation of the gene-disease association dataset" section).

		Examples	
S. No	**Types of redundancy**	**Disease terms**	**Modified term**
1	Target organ of disease	'Malignant neoplasm of ovary', 'ovarian neoplasm', 'Epithelial ovarian cancer'	Ovarian cancer
2	Age of onset of disease	'Adult type dermatomyositis', 'Dermatomyositis, Childhood Type','Dermatomyositis'	Dermatomyositis
3	Synonyms of disease	'Mental Depression', 'Major Depressive Disorder', 'Depressive disorder'	Mental Depression
4	Severity of disease	'Mental disorder', 'Mental disorder, severe', 'Mental disorder, acute', 'mental disorder, chronic'	Mental disorder

Table 1. Rules for redundancy elimination in gene-disease association dataset.

Four different algorithms have been used to predict the risk of comorbidity in women with PCOS. The comorbidity scores are illustrated as dynamic heat maps created using AnyChart JS[80] package.

Based on shared genes. This method is based on the principle that disease relationships are dependent on their shared genes[81]. A score to predict the risk of diseases D_i and D_j to co-occur is calculated using the below equation

$$Comorbidity_{sharedgenes}\left(D_i, D_j\right) = \left[\frac{\left(G_{D_i} \cap G_{D_j}\right)}{min\left(G_{D_i}, G_{D_j}\right)}\right] \times 100$$

where G_{D_i} and G_{D_j} are PCOS genes associated with diseases D_i and D_j.

The score is directly proportional to the number of shared genes; hence a higher score indicates a higher risk of comorbidity.

Based on the uniqueness of shared genes. This method is based on the observation that diseases, whose genes are not associated with multiple diseases, have a higher comorbidity risk as compared to diseases caused by genes associated with multiple diseases[82].

The uniqueness of ith gene 'g_i' associated with diseases $D_{i,}$ D_j is calculated as:

$$Uniqueness\left(g_i\right) = \left[1 - \sqrt{\frac{D_{g_i}}{D^T}}\right]$$

where D^T represents the total number of diseases in the gene-disease dataset and D_{g_i} is the number of diseases associated with ith gene.

If $ngenes \in D_i \cap D_j$ then, comorbidity of each disease pair is calculated as follows:

$$Comorbidity_{uniqueness}\left(D_i, D_j\right) = \sum_{n=1}^{n} \left[Uniqueness\left(g_i\right)\right]_n$$

The score is directly proportional to the number of uniquely shared genes, hence a higher score indicates a higher risk of comorbidity for the pair of diseases.

Based on the biological process and molecular function of associated genes. This algorithm is based on the inference that 95% of disease links can be predicted by the functional overlap of the associated genes[81]. Disease pair comorbidity risk is calculated and scored as per the standard Jaccard index[83].

$$Comorbidity_{ontology}\left(D_i, D_j\right) = \left[\left|\frac{GO_i \cap GO_j}{GO_i \cup GO_j}\right|\right] \times 100$$

where GO_i and GO_j are the set of distinct molecular functions and biological processes for genes of diseases i and j respectively as retrieved from Gene Ontology (GO) database.

The score is directly proportional to the functional overlap of disease-associated genes and therefore higher score indicates a higher risk of comorbidity for the pair of diseases.

Based on network separation of disease genes in the human interactome. Diseases whose genes are located closer in the human interactome have a higher probability of co-occurrence as compared to diseases with genes spread apart in the network[84]. Experimentally validated human protein–protein interactions from STRING v11[68] were used for the algorithm. The comorbidity score is calculated as:

$$Comorbidity_{Shortest\ path}\left(D_i, D_j\right) = D_{ij} - \frac{D_{ii} + D_{jj}}{2}$$

where D_{ii} and D_{jj} is the average of minimum distances of each gene associated with disease i and j respectively and D_{ij} is the average of minimum distances between genes of diseases i and j.

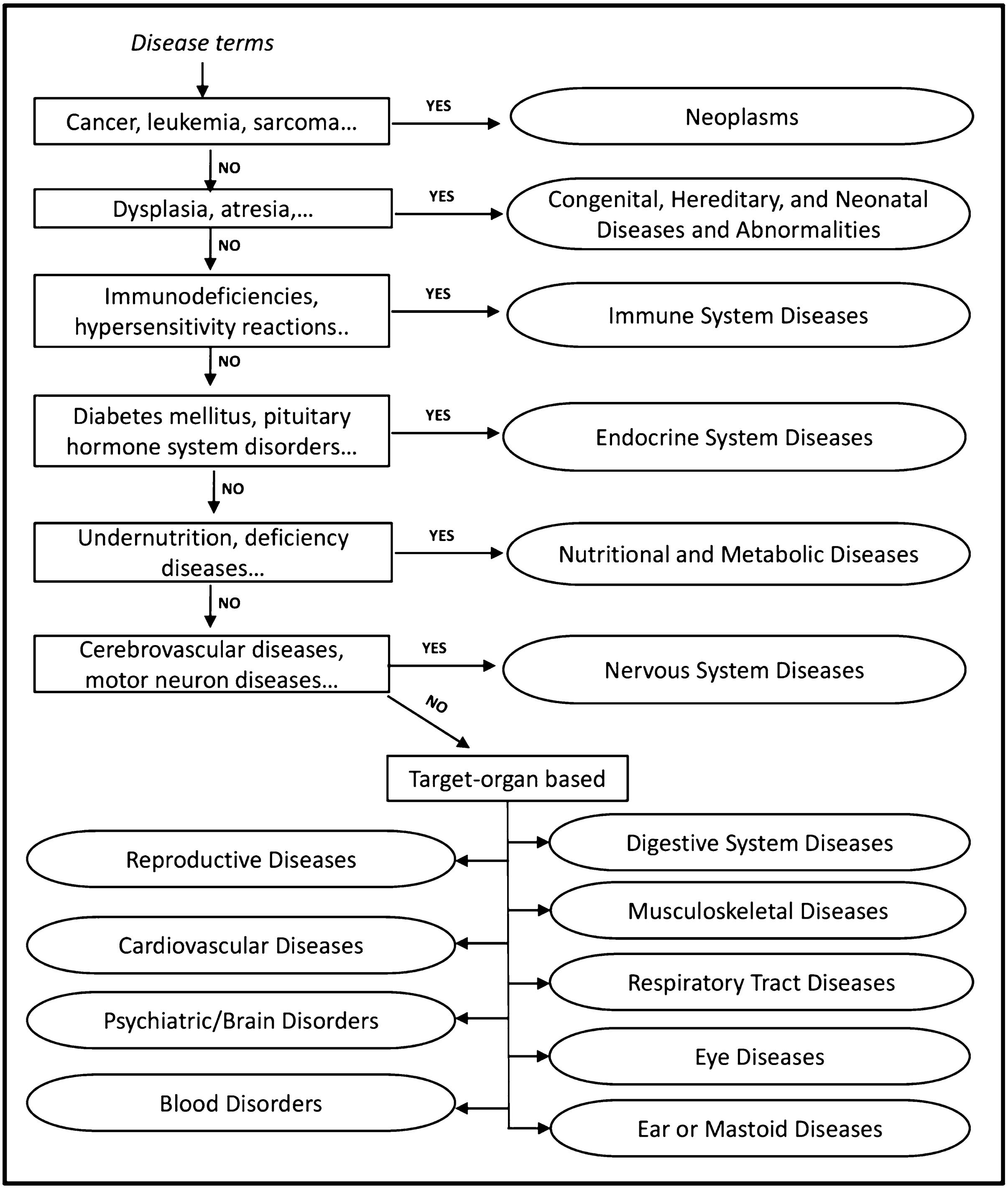

Disease terms
Cancer, leukemia, sarcoma…
YES
Neoplasms
NO
Dysplasia, atresia,…
YES
Congenital, Hereditary, and Neonatal Diseases and Abnormalities
NO
Immunodeficiencies, hypersensitivity reactions..
YES
Immune System Diseases
NO
Diabetes mellitus, pituitary hormone system disorders…
YES
Endocrine System Diseases
NO
Undernutrition, deficiency diseases…
YES
Nutritional and Metabolic Diseases
NO
Cerebrovascular diseases, motor neuron diseases…
YES
Nervous System Diseases
NO
Target-organ based
Reproductive Diseases
Digestive System Diseases
Musculoskeletal Diseases
Cardiovascular Diseases
Respiratory Tract Diseases
Psychiatric/Brain Disorders
Eye Diseases
Blood Disorders
Ear or Mastoid Diseases

Figure 4. ICD-11 based rules for non-redundant categorization of disease terms. Ovals represent retitled parent disease terms.

Since the score represents the network-based separation of disease-associated genes, a lower score indicates higher risk of comorbidity for the pair of diseases.

Network analysis. This tool can be used for visualization of disease networks, identification of enriched pathways, and prioritization of disease genes. Vis.js[85] visualization library was used for dynamic network creation and visualization. The tool has three modules as described below.

Disease-disease network. A dynamic subset of the human disease network[86] can be created for a selected group of diseases. Diseases are represented as nodes and the size of a node is proportional to the number of genes or miRNAs associated with the disease. Disease nodes are connected by edges based on the number of shared genes or miRNAs between them. Users can select multiple diseases for the identification of enriched pathways in these diseases.

Pathway enrichment analysis. The disease-pathway associations are inferred based on mapping disease-associated genes and target genes of associated miRNAs to their pathways[87]. Enriched pathways are identified based on hypergeometric distribution with the threshold p value set as 0.05 (gene dataset) and 0.001(miRNA dataset) based on the data size. Users can select pathways and visualize the network. Each pathway is represented as a node and is connected to other pathways in the network based on common genes or miRNAs. The thickness of the edge is proportional to the number of shared genes or miRNAs. If gene dataset is selected then, the enriched pathways can be examined for the identification of critical hub and bottleneck genes through the *Gene network analysis* module.

Gene network analysis. Experimentally validated interactions from STRING v11[68] were used for creating gene interaction networks for enriched pathways. Critical genes in these pathways were identified based on network topological properties such as degree, closeness centrality, and betweenness centrality calculated using graph package in R[88]. The hub and bottleneck genes were defined based on the study of Rakshit et al.[89].

Hub genes: *Degree* > (*Mean of Degree* + (2* *Standard Deviation*)) OR *Closeness centrality* > (*Mean of closeness centrality* + (2* *Standard Deviation*)).

Bottleneck genes: *Degree* < (*Mean of Degree*) AND *Betweenness centrality* > (*Mean of Betweenness centrality*).

Venn analysis. The common and unique list of genes, pathways, and ontologies can be identified for a selected list of diseases using this tool. jvenn source code [16] was used to develop the interactive 6-way Venn diagram.

Data availability

The data can be accessed and downloaded from the PCOSKB$_{R2}$ portal (http://www.pcoskb.bicnirrh.res.in/).

Received: 22 June 2020; Accepted: 17 August 2020
Published online: 07 September 2020

References

1. Wang, F. *et al.* Alternative splicing of the androgen receptor in polycystic ovary syndrome. *Proc. Natl. Acad. Sci. USA* **112**, 4743–4748 (2015).
2. Fauser, B. C. J. M. *et al.* Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome. *Hum. Reprod.* **19**, 41–47 (2004).
3. Azziz, R. Diagnostic criteria for polycystic ovary syndrome: a reappraisal. *Fertil. Steril.* https://doi.org/10.1016/j.fertnstert.2005.01.085 (2005).
4. Azziz, R. *et al.* The androgen excess and PCOS Society criteria for the polycystic ovary syndrome: the complete task force report. *Fertil. Steril.* **91**, 456–488 (2009).
5. Yedulapuram, S. H., Gunda, M., Moola, N. R. & Kadarla, R. K. An overview on polycystic ovarian syndrome. *Asian J. Pharm. Res. Dev.* **7**, 72–80 (1970).
6. Gilbert, E. W., Tay, C. T., Hiam, D. S., Teede, H. J. & Moran, L. J. Comorbidities and complications of polycystic ovary syndrome: an overview of systematic reviews. *Clin. Endocrinol.* **89**, 683–699 (2018).
7. Kazemi, M. *et al.* Comprehensive evaluation of type 2 diabetes and cardiovascular disease risk profiles in reproductive-age women with polycystic ovary syndrome: a large canadian cohort. *J. Obstet. Gynaecol. Canada* **41**, 1453–1460 (2019).
8. Kakoly, N. S., Moran, L. J., Teede, H. J. & Joham, A. E. Cardiometabolic risks in PCOS: a review of the current state of knowledge. *Exp. Rev. Endocrinol. Metab.* **14**, 23–33 (2019).
9. Dokras, A., Clifton, S., Futterweit, W. & Wild, R. Increased prevalence of anxiety symptoms in women with polycystic ovary syndrome: Systematic review and meta-analysis. *Fertil. Steril.* **97**, 225-230.e2 (2012).
10. Chen, S. F., Yang, Y. C., Hsu, C. Y. & Shen, Y. C. Risk of bipolar disorder in patients with polycystic ovary syndrome: a nationwide population-based cohort study. *J. Affect. Disord.* **263**, 458–462 (2020).
11. Thannickal, A. *et al.* Eating, sleeping and sexual function disorders in women with polycystic ovary syndrome (PCOS): a systematic review and meta-analysis. *Clin. Endocrinol. (Oxf)* **92**, 338–349 (2020).
12. Barthelmess, E. K. & Naz, R. K. Polycystic ovary syndrome: current status and future perspective. *Frontiers Biosci. Elite* **6E**, 104–119 (2014).
13. Joseph, S., Barai, R. S., Bhujbalrao, R. & Idicula-Thomas, S. PCOSKB: A knowledgebase on genes, diseases, ontology terms and biochemical pathways associated with polycystic ovary syndrome. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkv1146 (2016).
14. Afiqah-Aleng, N., Harun, S., A-Rahman, M. R. A., Nor Muhammad, N. A. & Mohamed-Hussein, Z. A. PCOSBase: a manually curated database of polycystic ovarian syndrome. *Database* https://doi.org/10.1093/database/bax098 (2017).
15. Maniraja, J. M., Vetrivel, U., Munuswamy, D. & Melanathuru, V. PCOSDB: PolyCystic ovary syndrome DataBase for manually curated genes associated with the disease. *Bioinformation* **12**, 4–8 (2016).
16. Bardou, P., Mariette, J., Escudié, F., Djemiel, C. & Klopp, C. Jvenn: an interactive Venn diagram viewer. *BMC Bioinform.* **15**, 293 (2014).
17. Scicchitano, P. *et al.* Cardiovascular risk in women with PCOS. *Int. J. Endocrinol. Metab.* **10**, 611–618 (2012).
18. Kulshreshtha, B., Singh, S. & Arora, A. Family background of Diabetes Mellitus, obesity and hypertension affects the phenotype and first symptom of patients with PCOS. *Gynecol. Endocrinol.* **29**, 1040–1044 (2013).
19. Sirmans, S. M., Parish, R. C., Blake, S. & Wang, X. Epidemiology and comorbidities of polycystic ovary syndrome in an indigent population. *J. Investig. Med.* **62**, 868–874 (2014).
20. Elting, M. W., Korsen, T. J., Bezemer, P. D. & Schoemaker, J. Prevalence of diabetes mellitus, hypertension and cardiac complaints in a follow-up study of a Dutch PCOS population. *Hum. Reprod.* **16**, 556–560 (2001).
21. Rasgon, N. L. *et al.* Depression in women with polycystic ovary syndrome: clinical and biochemical correlates. *J. Affect. Disord.* **74**, 299–304 (2003).
22. Rodriguez-Paris, D. *et al.* Psychiatric disorders in women with polycystic ovary syndrome. *Psychiatr. Pol.* **53**, 955–966 (2019).

23. Rassi, A. *et al.* Prevalence of psychiatric disorders in patients with polycystic ovary syndrome. *Compr. Psychiatry* **51**, 599–602 (2010).
24. Hung, J. H. *et al.* Risk of psychiatric disorders following polycystic ovary syndrome: a nationwide population-based cohort study. *PLoS One* **9**, e97041 (2014).
25. Annagür, B. B., Kerimoglu, ÖS., Tazegül, A., Gündüz, Ş & Gençoglu, B. B. Psychiatric comorbidity in women with polycystic ovary syndrome. *J. Obstet. Gynaecol. Res.* **41**, 1229–1233 (2015).
26. Brutocao, C. *et al.* Psychiatric disorders in women with polycystic ovary syndrome: a systematic review and meta-analysis. *Endocrine* **62**, 318–325 (2018).
27. Cheung, B. M. Y. The hypertension-diabetes continuum. *J. Cardiovasc. Pharmacol.* **55**, 333–339 (2010).
28. Zhao, L. *et al.* Estrogen receptor 1 gene polymorphisms are associated with metabolic syndrome in postmenopausal women in China 11 Medical and Health Sciences 1103 Clinical Sciences. *BMC Endocr. Disord.* **18**, 65 (2018).
29. Jiao, X. *et al.* Variant alleles of the ESR1, PPARG, HMGA2, and MTHFR genes are associated with polycystic ovary syndrome risk in a Chinese population: A case-control study. *Front. Endocrinol. (Lausanne)* **9**, 504 (2018).
30. Jakimiuk, A. J., Weitsman, S. R., Yen, H. W., Bogusiewicz, M. & Magoffin, D. A. Estrogen receptor α and β expression in theca and granulosa cells from women with polycystic ovary syndrome. *J. Clin. Endocrinol. Metab.* **87**, 5532–5538 (2002).
31. Wong, W. T., Tian, X. Y. & Huang, Y. Endothelial dysfunction in diabetes and hypertension: cross talk in RAS, BMP4, and ROS-dependent COX-2-derived prostanoids. *J. Cardiovasc. Pharmacol.* **61**, 204–214 (2013).
32. Schmidt, J. *et al.* Differential expression of inflammation-related genes in the ovarian stroma and granulosa cells of PCOS women. *Mol. Hum. Reprod.* **20**, 49–58 (2014).
33. Supriya, R. *et al.* Adipokines demonstrate the interacting influence of central obesity with other cardiometabolic risk factors of metabolic syndrome in Hong Kong Chinese adults. *PLoS One* **13**, e0201585 (2018).
34. Gacka, M. & Adamiec, R. Mutations of peroxisome proliferator-activated receptor gamma (PPARgamma): clinical implications. *Postepy Hig. Med. Dosw. (Online)* **58**, 483–489 (2004).
35. Tsatsakis, A. M., Zafiropoulos, A., Tzatzarakis, M. N., Tzanakakis, G. N. & Kafatos, A. Relation of PON1 and CYP1A1 genetic polymorphisms to clinical findings in a cross-sectional study of a Greek rural population professionally exposed to pesticides. *Toxicol. Lett.* **186**, 66–72 (2009).
36. Wang, Y. *et al.* Evidence for association between paraoxonase 1 gene polymorphisms and polycystic ovarian syndrome in southwest Chinese women. *Eur. J. Endocrinol.* **166**, 877–885 (2012).
37. Ohashi, K., Ouchi, N. & Matsuzawa, Y. Adiponectin and Hypertension. *Am. J. Hypertens.* **24**, 263–269 (2011).
38. Davis, S. K. *et al.* Association of adiponectin with type 2 diabetes and hypertension in African American men and women: The Jackson Heart Study. *BMC Cardiovasc. Disord.* **15**, 13 (2015).
39. Mirza, S. S., Shafique, K., Shaikh, A. R., Khan, N. A. & Anwar Qureshi, M. Association between circulating adiponectin levels and polycystic ovarian syndrome. *J. Ovarian Res.* **7**, 18 (2014).
40. Cooney, L. G. & Dokras, A. Depression and anxiety in polycystic ovary syndrome: etiology and treatment. *Current Psychiatry Rep.* **19**, 83 (2017).
41. Meczekalski, B., Pérez-Roncero, G. R., López-Baena, M. T., Chedraui, P. & Pérez-López, F. R. The polycystic ovary syndrome and gynecological cancer risk. *Gynecol. Endocrinol.* **36**, 289–293 (2020).
42. Feng, Y. *et al.* Effects of androgen and leptin on behavioral and cellular responses in female rats. *Horm. Behav.* **60**, 427–438 (2011).
43. Kamalanathan, S., Sahoo, J. & Sathyapalan, T. Pregnancy in polycystic ovary syndrome. *Indian J. Endocrinol. Metab.* **17**, 37 (2013).
44. Hartanti, M. D. *et al.* Could perturbed fetal development of the ovary contribute to the development of polycystic ovary syndrome in later life?. *PLoS One* **15**, e0229351 (2020).
45. Scalici, E. *et al.* Circulating microRNAs in follicular fluid, powerful tools to explore in vitro fertilization process. *Sci. Rep.* **6**, 1–10 (2016).
46. He, T. *et al.* MicroRNA-141 and MicroRNA-200c are overexpressed in granulosa cells of polycystic ovary syndrome patients. *Front. Med.* **5**, 299 (2018).
47. Tesfaye, D. *et al.* Potential role of microRNAs in mammalian female fertility. *Reprod. Fertil. Dev.* **29**, 8–23 (2017).
48. Lykoudi, A. *et al.* Dysregulated placental microRNAs in early and Late onset Preeclampsia. *Placenta* **61**, 24–32 (2018).
49. Lu, J., Wang, Z., Cao, J., Chen, Y. & Dong, Y. A novel and compact review on the role of oxidative stress in female reproduction. *Reprod. Biol. Endocrinol.* **16**, 80 (2018).
50. Lundberg, F. E., Iliadou, A. N., Rodriguez-Wallberg, K., Gemzell-Danielsson, K. & Johansson, A. L. V. The risk of breast and gynecological cancer in women with a diagnosis of infertility: a nationwide population-based study. *Eur. J. Epidemiol.* **34**, 499–507 (2019).
51. Yin, W., Falconer, H., Yin, L., Xu, L. & Ye, W. Association between polycystic ovary syndrome and cancer risk. *JAMA Oncol.* **5**, 106–107 (2019).
52. Wolf, W. M., Wattick, R. A., Kinkade, O. N. & Olfert, M. D. Geographical prevalence of polycystic ovary syndrome as determined by region and race/ethnicity. *Int. J. Environ. Res. Public Health* **15**, 2589 (2018).
53. Khan, M. J., Ullah, A. & Basit, S. Genetic basis of polycystic ovary syndrome (PCOS): current perspectives. *Appl. Clin. Genet.* **12**, 249–260 (2019).
54. Choudhary, A., Jain, S. & Chaudhari, P. Prevalence and symptomatology of polycystic ovarian syndrome in Indian women: is there a rising incidence?. *Int. J. Reprod. Contracept. Obstet. Gynecol.* **6**, 4971 (2017).
55. Belenkaia, L. V., Lazareva, L. M., Walker, W., Lizneva, D. V. & Suturina, L. V. Criteria, phenotypes and prevalence of polycystic ovary syndrome. *Minerva Ginecol.* **71**, 211–225 (2019).
56. Nouraldein, M., Hamad, M., Abdelgadir, M. A., Omer, M. & Hussein, M. Prevalence of Stein-Leventhal syndrome. *EC EC Gynaecol.* **9**, 52–55 (2020).
57. Canese, K. & Weis, S. PubMed: The bibliographic database. *NCBI Handb.* (2013)
58. Lipscomb, C. E. Medical Subject headings (MeSH). *Bull. Med. Libr. Assoc.* **88**, 265–266 (2000).
59. Brown, G. R. *et al.* Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.* **43**, D36-42 (2015).
60. Sherry, S. T. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* https://doi.org/10.1093/nar/29.1.308 (2001).
61. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkx1098 (2018).
62. Bateman, A. *et al.* UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204–D212 (2015).
63. Rose, P. W. *et al.* The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkw1000 (2017).
64. Gene Ontology Consortium. Gene ontology consortium: going forward. *Nucleic Acids Res.* **43**, D1049–D1056 (2015).
65. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109–D114 (2012).
66. Hamosh, A., Scott, A. F., Amberger, J., Valle, D. & McKusick, V. A. Online mendelian inheritance in man (OMIM). *Hum. Mutat.* **15**, 57–61 (2000).
67. Fabregat, A. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkx1132 (2018).
68. Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gky1131 (2019).

69. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkz1021 (2020).
70. NIH, N. MedGen. *NIH* (2016).
71. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkt1113 (2014).
72. Davis, A. P. *et al.* The comparative toxicogenomics database: update 2019. *Nucleic Acids Res.* **47**, D948–D954 (2019).
73. England, G. The 100,000 genomes project protocol v3 genomics England. *Genomics Engl. Protoc.* https://doi.org/10.6084/m9.figshare.4530893.v2 (2017).
74. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkt1229 (2014).
75. Flint, J. GWAS. *Curr. Biol.* https://doi.org/10.1016/j.cub.2013.01.040 (2013).
76. Bravo, À, Piñero, J., Queralt-Rosinach, N., Rautschka, M. & Furlong, L. I. Extraction of relations between genes and diseases from text and large-scale data analysis: Implications for translational research. *BMC Bioinform.* **16**, 55 (2015).
77. Bundschus, M., Dejori, M., Stetter, M., Tresp, V. & Kriegel, H. P. Extraction of semantic biomedical relations from text using conditional random fields. *BMC Bioinform.* **9**, 207 (2008).
78. Rani, J., Shah, A. R. & Ramachandran, S. pubmed.mineR: An R package with text-mining algorithms to analyse PubMed abstracts. *J. Biosci.* **40**, 671–682 (2015).
79. The Lancet. *ICD-11. The Lancet* (Elsevier, Asterdam, 2019). https://doi.org/10.1016/S0140-6736(19)31205-X.
80. Heat Map Chart|Basic Charts|AnyChart Documentation. Available at: https://docs.anychart.com/Basic_Charts/Heat_Map_Chart . (Accessed: 5th May 2020)
81. Rubio-Perez, C. *et al.* Genetic and functional characterization of disease associations explains comorbidity. *Sci. Rep.* **7**, 1–14 (2017).
82. Carson, M. B., Liu, C., Lu, Y., Jia, C. & Lu, H. A disease similarity matrix based on the uniqueness of shared genes. *BMC Med. Genom.* **10**, 26 (2017).
83. Sun, K., Gonçalves, J. P., Larminie, C. & Pržulj, N. Predicting disease associations via biological network analysis. *BMC Bioinform.* **15**, 304 (2014).
84. Menche, J. *et al.* Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**(80), 347–841 (2015).
85. vis.js. Available at: https://visjs.org/. (Accessed: 9th May 2020)
86. Goh, K. . Il. *et al.* The human disease network. *Proc. Natl. Acad. Sci. USA* **104**, 8685–8690 (2007).
87. miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database - PubMed. Available at: https://pubmed.ncbi.nlm.nih.gov/31647101/. (Accessed: 4th August 2020)
88. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJ. Complex Syst.* **1695**(5), 1–9 (2006).
89. Rakshit, H., Rathi, N. & Roy, D. Construction and analysis of the protein-protein interaction networks based on gene expression profiles of Parkinson's Disease. *PLoS ONE* **9**, e103047 (2014).

Acknowledgements

We are grateful to Ms. Ulka Gawde and Mr. Viddyansh Srivastava for assisting in the statistical calculations; Ms. Alifiya Sadriwala for literature review; Ms. Karishma Desai, Mr. Tanmay Shirsat and Ms. Shuvechha Mukherjee for reviewing the manuscript and database. This work [RA/891/03-2020] was supported by grants received from the Department of Biotechnology [BT/PR26350/BID/7/810/2017], Department of Health Research [R.12020/01/2019-HR] and Indian Council of Medical Research.

Author contributions

S.I-T. planned, designed, and supervised the study. S.B., M.S., and I.K. curated and annotated the data. R.S.B., I.K., and K.P. developed the interface and algorithms for analysis. M.S., I.K., and S.I-T. wrote the manuscript. All authors approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-71418-8.

Correspondence and requests for materials should be addressed to S.I.-T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.