An integrative biomedical informatics approach to elucidate the similarities between pre-eclampsia and hypertension

Guillermo Lopez-Campos[a], Emma Bonner[a], Lana McClements[a,b]

[a] Centre for Experimental Medicine, Queen's University of Belfast, Belfast, Northern Ireland, United Kingdom
[b] School of Life Sciences, University of Technology Sydney, PO Box 123 Broadway, New South Wales, Australia

Abstract

Pre-eclampsia is a pregnancy condition affecting 5-10% of pregnancies, and it is the leading cause of death in pregnanc,y associated with increased risk of cardiovascular disease later in life. Despite research, the pathogenesis of pre-eclampsia is still poorly understood. In this paper, we investigate the overlapping pathogenic mechanisms between pre-eclampsia and adult hypertension using an integrative biomedical informatics strategy that combined text mining techniques, to identify genes and proteins, with geneset analyses, generating knowledge on the pathways and mechanisms involved in these conditions. We identified several overlapping pathogenic pathways/systems including metabolic pathways, developmental biology pathways, immune system, haemostasis, tyrosine kinase pathways, extracellular matrix and oxidative stress pathways. This bioinformatics approach could be applied for investigating mechanistic pathways of other disorders.

Keywords:

Preeclampsia, Computational Biology, Text Mining.

Introduction

Pre-eclampsia is a condition affecting 5-10% of pregnancies. It is clinically characterised by the new onset of hypertension and proteinuria or other organ damage post 20 weeks gestation [1]. There are different types of pre-eclampsia diagnosed according to the time of onset in pregnancy i.e. early (before 34 weeks gestation) and late (after 34 weeks gestations) or according to the severity of the symptoms[2]. Pre-eclampsia is a leading cause of mortality and morbidity, claiming approximately 80,000 maternal and 500,000 foetal deaths every year[3]. Long term complications associated with pre-eclampsia include type 2 diabetes mellitus and cardiovascular disease (CVD) for both mothers and adult offspring, later in life[4,5]

During the early stages of pregnancy, a group of differentiated cells forming blastocysts, become implanted in the lining of the uterus. The outer layer of the blastocyst consists of foetal trophoblast cells which invade the spiral uterine arteries after implantation[6]. The invasion of trophoblast cells into the lining of the uterus results in remodelling of spiral uterine arteries to allow an unlimited supply of oxygen and nutrients to the foetus. These changes involve the replacement of maternal endothelial cells with invasive trophoblasts, which develop to form a large section of the placenta, leading to irreversible dilation of blood vessels and loss of elastic tissue, therefore preventing restriction to blood flow and vasomotor control which regulate blood pressure[7]. Impaired trophoblast invasion has been implicated in the pathogenesis of pre-eclampsia. This results in the absence of uteroplacental remodelling of spiral arteries leading to placental ischaemia and causing insufficient perfusion and delivery of the oxygen and nutrients to the foetus which can affect the foetal growth[8]. In an attempt to overcome the lack of blood and nutrient supply to the foetus, maternal blood pressure becomes gradually elevated. The restricted blood supply to the placenta causes the release of chemical mediators and hormones into the maternal circulation, altering endothelial cell activity and resulting in endothelial cell dysfunction. This suppresses the release of vasodilatory factors such as nitric oxide (NO) and increases the production of vasoconstrictors while inhibiting anticoagulant synthesis and increasing procoagulant production[9]. Dysbalance of vasodilatory and vasoconstrictive factors increases the risk of blood clotting and can lead to an increase in blood pressure. Endothelial dysfunction has been implicated in the pathogenesis of pre-eclampsia[10,11], however endothelial dysfunction following pregnancy complicated by severe pre-eclampsia appears to also persist 10-20 years after pregnancy[12].

As mentioned above, women and children born as a result of pre-eclamptic pregnancies have increased risk of CVD later in life, particularly if pre-eclampsia was severe (blood pressure ≥160/110 mmHg, thrombocytopenia, impaired liver function, progressive renal insufficiency, pulmonary oedema, and cerebral or visual disturbances). Based on a meta-analysis of prospective and retrospective cohort studies including 198,252 women with pre-eclampsia, the relative risks of developing hypertension, ischaemic heart disease and stroke were 3.70 after 14.1 years, 2.16 after 11.7 years, 1.81 after 10.4 years, respectively[13]. Furthermore, children born to mothers who suffered pre-eclampsia in pregnancy were also at increased risk of high BP and stroke in later life[14].

Pre-eclampsia and cardiovascular disease are affected by similar risk factors such as hypertension, obesity, insulin resistance, diabetes, a family history of CVD and genotypes related to CVD, suggesting that similar pathogenic mechanisms could be present in both diseases[12,15]. Biomarker discovery and characterisation is an important area of research both in pre-eclampsia and CVD that led to the identification of a number of important markers. However, despite research in this field, there is still a lack of knowledge in relation to the mechanisms underpinning the pathogenesis of pre-eclampsia and future increased risk of CVD.

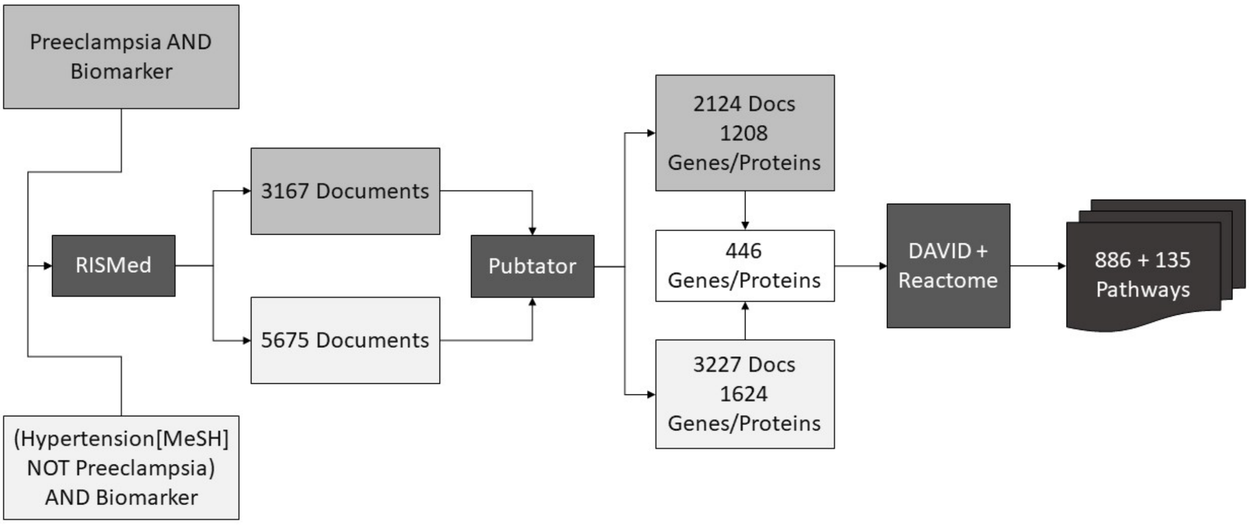
Preeclampsia AND Biomarker
RISMed
(Hypertension[MeSH] NOT Preeclampsia) AND Biomarker
3167 Documents
5675 Documents
Pubtator
2124 Docs 1208 Genes/Proteins
446 Genes/Proteins
3227 Docs 1624 Genes/Proteins
DAVID + Reactome
886 + 135 Pathways

Figure 1 – Overview of the methodology and the results from the the query terms to the final and total number (sum of Reactome and DAVID results) potentially shared pathways and genesets between preeclampsia and hypertension

Scientific literature presents the knowledge generated in research and practice and it is a continuously growing wealth of data that is enriched on a daily base with new publications. These data represent an opportunity for the development of research strategies aiming to generate new biomedical knoweldge that is hidden in this vast amount of information, or the generation of new research hyptheses. Therefore, it has attracted the interest of biomedical informatics fostering the development of methods and tools designed to mine it. An important element of these methodologies has been focused on the identification of diseases, genes or proteins cited in scientific texts and the identification of interaction networks. Biomedical relation extraction is an approach that has shown precedence in similar studies providing cross-sectional and domain-specific views of biomedical research literature [16–18].

In this study, we endeavoured to investigate an *"in silico"* approach, which uses the information embedded in the literature to integrate relevant biomarkers identified in pre-eclampsia and hypertension and contextualise these in terms of significant biological pathways which are common denominators in both conditions. The approach presented here provides a mechanistic interpretation of the commonalities between pre-eclampsia and hypertension, but it could be applied to study pathogenesis of any other disorder.

Methods

The methodology applied in this work consisted of three sequential stages including: i) document search and retrieval, ii) document annotation and gene/protein extraction and iii) mechanistic analysis. These steps are further detailed below.
The first step in our analyses included the definition of the relevant Pubmed queries, which were used to identify the relevant elements in the literature associated with the two conditions of interest for this study.

- Preeclampsia query was built as "("pre-eclampsia"[MeSH Terms] OR "pre-eclampsia"[All Fields] OR "preeclampsia"[All Fields]) AND ("biomarkers"[MeSH Terms] OR "biomarkers"[All Fields] OR "biomarker"[All Fields])". This query aimed to retrieve a broad range of elements that could be associated with this condition.
- Hypertension query was built as "(Hypertension[MeSH] NOT ("pre-eclampsia"[MeSH Terms] OR "pre-eclampsia"[All Fields] OR "preeclampsia"[All Fields])) AND Biomarker". This query focused on the retrieval of documents annotated as hypertension as a MeSH term while excluding preeclampsia.

We used the R package RISMed (2.1.7) designed to retrieve and download contents from the NCBI databases into R. For this purpose both of these queries were then passed as arguments to the function "EUtilsSummary" to identify the relevant documents in Pubmed that were subsequently downloaded using the "EUtilsGet" function. This resulted in two R data frames containing information associated with these documents and in particular the relevant Pubmed IDs that were used as input in the second stage for annotation of the literature.

The second stage of the analysis consisted of the annotation of the biological terms of relevance (genes/proteins); for this purpose, rather than developing a new methodology, we relied on the annotations provided by Pubtator[19]. Pubtator annotates Pubmed documents using different "named-entity-recognition" algorithms around four bioconcepts Gene; Chemical; Disease; Species. Although Pubtator was initially developed as a web service at the NCBI ftp site we used the downloadable version of the results from its annotations.

The third and the final stage consisted of the mechanistic analysis of the identified biomarkers. In this analyses we selected two knowledge bases to analyse the enrichment in certain biological processes or functions: DAVID [20], broad and containing multiple annotations such as gene ontology terms, Uniprot keywords or KEGG terms for multiple organisms; and Reactome [21], a highly curated database of biological reactions in humans. Reactome identifies the enrichment in its contents using an overrepresentation analysis based in the hypergeometric test and using a FDR (Benjamini-Hochberg) p-value correction to adjust the significance for multiple comparisons. DAVID uses EASE, a modified version of the hypergeometric test and allows the selection of multiple p-value correction methods. In both cases the input provided was a list of NCBI Gene IDs that were translated into their own internal IDs to match their contents.

Results

An overall overview of the methodology developed and the results obtained are presented in Figure 1.

The literature search allowed us to identify and retrieve almost 9000 documents in total, 3167 for preeclampsia and 5675 for hypertension. As a quality control we compared the PMID between these two datasets to ensure that there was no overlap and that the results would not be misled by publications present in both datasets.

The annotation of these two sets of pubmed documents allowed us to identify approximately 5000 different documents that were annotated with at least one gene or protein providing 2124 in the pre-eclampsia dataset and 3227 in the hypertension dataset. These documents contained 2386 genes/proteins of which 446 were present in both datasets (19% of the overall gene list and 37% of the genes identified in the pre-eclampsia dataset; Figure 2).

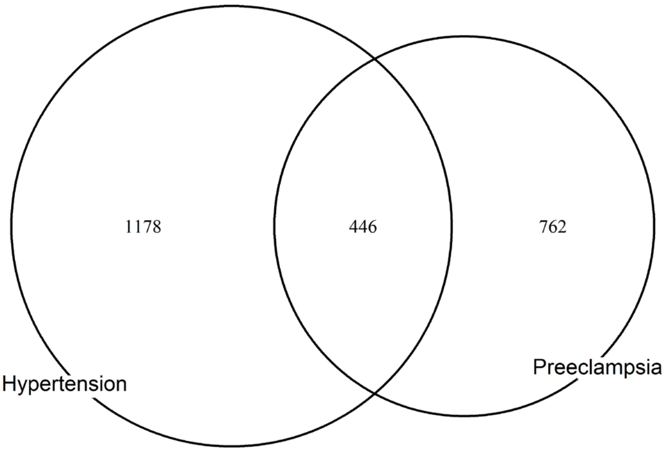

Figure 2 – Number of genes/proteins identified for each dataset and the set of 446 shared genes proteins that were used in the geneset analyses

Once we identified the relevant gene lists the next step was to perform the geneset analyses of the identified genes. As the significance is determined by the number of elements in the genelist we initially run three different analyses using Reactome with the following three subsets: 1- All genes (1624) identified for hypertension (HT.a); 2- All genes (1208) identified for preeclampsia (PE.a); 3- the 446 genes present in both hypertension and preeclampsia (Both). The significance threshold for enriched pathways was set as an adjusted p-value of less than 0.05. The number of significant pathways identified in each of these analyses are presented in the Table 1 and the shared pathways across these comparisons are visualised in Figure 3.

Table 1 – Number of Reactome pathways significantly enriched for each of the three genesets analysed.

Gene Subset (Number of Genes/proteins) (code)	Number of Pathways	Unique Pathways
All genes in Hypertension (1624) (HT.a)	20	7
All genes in Preeclampsia (1208) (PE.a)	107	21
Genes present in both Preeclampsia and Hypertension (446) (Both)	129	42

The set containing the genes/proteins present in both conditions ("Both" set) was the most enriched one with 129 significant pathways (Figure 4) whereas the hypertension set, which was the largest, was the least enriched with 20 different reactome pathways identified. An overlap analysis of these results allowed us to identfy 12 pathways in common for the three different genesets (Table 2).

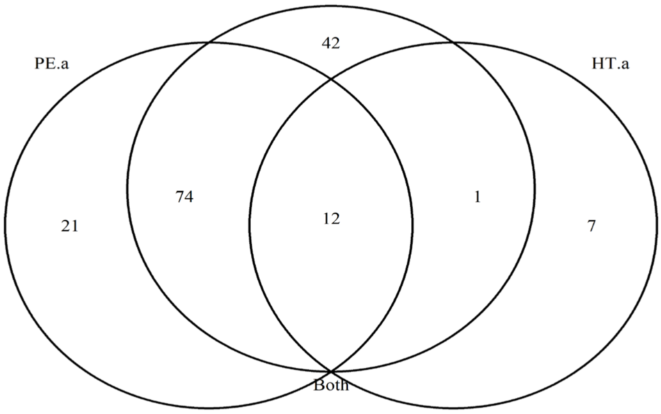

Figure 3 – Pathway enrichment for the three analysed sets, hypertension (HT.a), preeclampsia (PE.a) and the intersection between hypertension and preeclampsia (both)

Table 2 – List of the 12 statistically significant (adjusted p-value <0.05) pathways identified in Reactome for the three genesets analysed.

Pathway Name.
Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs)
Platelet degranulation
Response to elevated platelet cytosolic Ca2+
Post-translational protein phosphorylation
Platelet activation, signaling and aggregation
Integrin cell surface interactions
Chemokine receptors bind chemokines
Peptide ligand-binding receptors
Interleukin-10 signaling
Syndecan interactions
Formation of Fibrin Clot (Clotting Cascade)
PI5P, PP2A and IER3 Regulate PI3K/AKT Signaling

As different resources contain different information and annotations for the different genes we submitted the "both" (shared) dataset to DAVID for a complementary analysis. In this case we limit the query to the 446 shared elements between both pathologies. And set a significance threshold of benjamini adjusted p-value < 0.05. This resulted in the detection of significant enrichment in 886 different genesets made of pathways, gene ontology terms (Table 3) and other annotations.

Table 3 – List of the top 12 statistically significant (adjusted p-value <0.05) Gene Ontology Biologial Process terms identifed in DAVID for the intersected gene list

	Pathway Name		Pathway Name
1	Inflammatory response	7	Response to drug
2	Response to hypoxia	8	Platelet degranulation
3	Leukocyte migration	9	Aging

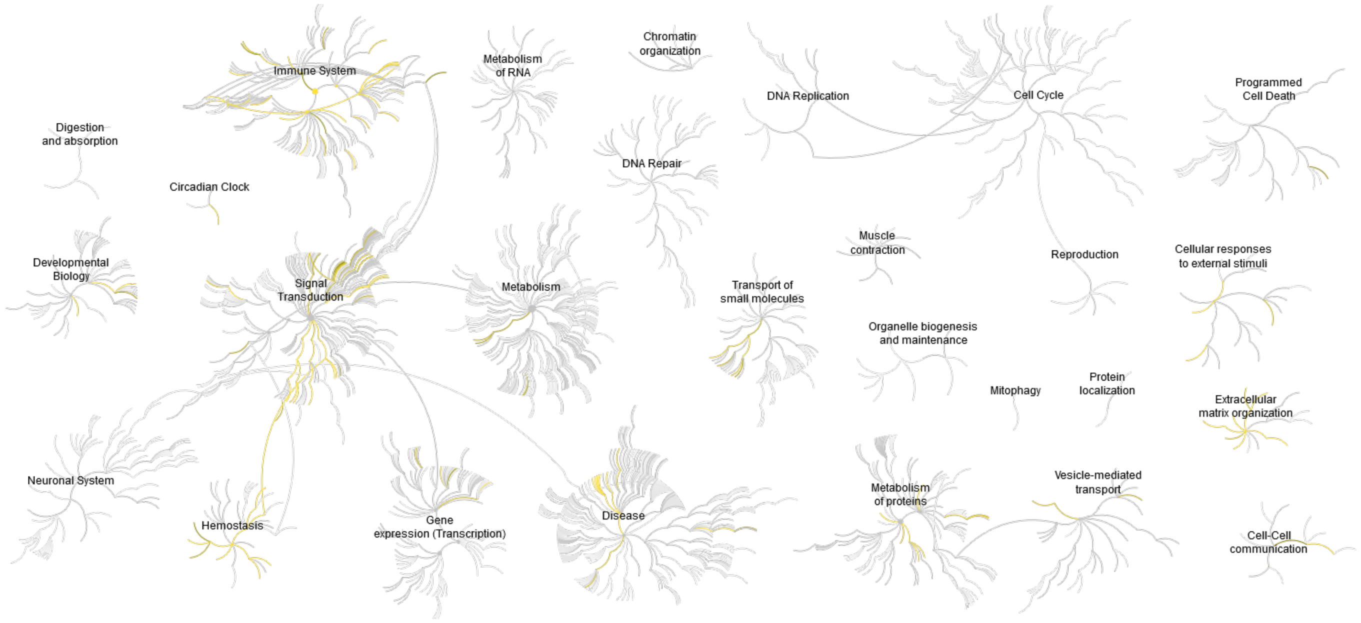

Figure 4 – Results from the Reactome analysis using the shared elements between pre-eclampsia and hypertension. The statistically significant elements are highlighted in yellow in the image.

4	Positive regulation of nitric oxide biosynthetic process	10	Positive regulation of smooth muscle cell proliferation
5	Positive regulation of gene expression	11	Positive regulation of angiogenesis
6	Angiogenesis	12	Regulation of blood pressure

Discussion

Using an integrated strategy we have been able to retrieve biomarkers (gene/proteins) that are shared between pre-eclampsia and hypertension. Our strategy allowed us to also identify genes from other organisms commonly used to model human diseases such as rats and mice. Although it would have been possible to map those genes to their human orthologues we decided not to do this therefore keeping our focus just on human biomarkers.

The identified genes were transformed into gene lists that were analysed using two different resources. As we mentioned before we were focused on human disease or models and, for this reason, we chose to run initially a more restrictive search using Reactome, as this is a highly curated database used to generate results from high-quality and restricted sources facilitating their biological interpretation. Overrepresentation analyses, similar to what we carried out here, are dependent on the size of the query lists and the size of the background available in the resource they are compared with. Therefore, finding differences in the composition and number of significant pathways identified between the lists containing the whole set of biomarkers for each condition and their intersection, was not completely unexpected. However, it was surprising that the smallest list containing the conserved biomarkers was the one showing the highest enrichment. From a biological perspective, this list, derived from the intersection of biomarkers identified in both pathologies, is expected to contain genes/proteins positively associated with the diseases. These genes would therefore be more biologically meaningful and would better capture the common underlying biological processes in both conditions.

The identified pathways for the shared set of genes are related to metabolism, developmental biology, immune system, haemostasis, tyrosine kinase, extracellular matrix and oxidative stress signalling. All these aspects provide coherent and meaningful biological information in the context of these pathologies. Interestingly we have identified using DAVID analyses, the GO terms associated with “Regulation of blood pressure” showing up as statistically significant and highly ranked. This is an expected result as both conditions show changes in blood pressure which validates our methodology as both conditions share a high blood pressure phenotype. In relation to the other enriched pathways identified, for example it is well recognised that inflammation (associated with the immune system) plays a significant role in hypertension[22,23]. Furthermore, pre-eclampsia is a pregnancy related condition therefore developmental biology plays a key role[6, 7].

Limitations and future work

The limitations of this study include some aspects of methodology which assume a perfect gene/protein identification in the annotation process not taking into the account the effects of potential false positive and negative biomarker hits. The second limitation is associated with the assumption of treating equally all the terms identified and not considering potential negations contained in the abstracts. An example of this is that we are including a gene into our analysis which could appear in the text as “gene XXX does not have an effect in preeclampsia”. However we tried to minimise this by basing our approach on relevant pathways and disease mechanisms.

As part of the future work, we are planning to address specifically the effects and quantity of false positives and negatives potentially present in the current version and we will try to develop new approaches which can help us identify and successfully manage negations found in the literature used as input.

In addition to this, as part of the future work we plan to develop an R package that will seamlessly integrate the methodology and processes that have been described in this work.

Finally, we plan to biologically validate some of the findings derived from this work.

Conclusions

There is a wealth of data and inforamtion available in public repositories that could be mined and analysed to generate new knowledge and research hypotheses. In this work, we combined different biomedical informatics tools, to retrieve and identify biomarkers in hypertension and pre-eclampsia and then analysed these to identifiy common underlying molecular mechanisms linking these two conditions.

This data-driven approach enabled us to identify the specific aspects and reactions associated with metabolic pathways, developmental biology, immune system, haemostasis, tyrosine kinase pathways, extracellular matrix and oxidative stress pathways as the most prominently involved in the pathogenesis of these two important conditions.

Our bioinformatics approach described in this paper is therefore applicable to any other similar diseases for the purpose of identifying overlapping or individual pathogenic mechanisms.

References

[1] American College of Obstetricians and Gynecologists, Task Force on Hypertension in Pregnancy. Hypertension in Pregnancy. Obstet Gynecol 2013;122:1122–1131. doi:10.1097/01.AOG.0000437382.03963.88.[2] J.M. Roberts, G.D. Pearson, J.A. Cutler, and M.D. Lindheimer, Summary of the NHLBI Working Group on Research on Hypertension During Pregnancy., *Hypertens. Pregnancy*. **22** (2003) 109–27. doi:10.1081/PRG-120016792.

[3] Khan KS, Wojdyla D, Say L, Gülmezoglu AM, Van Look PF. WHO analysis of causes of maternal death: a systematic review. Lancet. 2006;367:1066–1074 doi: 10.1016/S0140-6736(06)68397-9

[4] G.T.R. Manten, M.J. Sikkema, H.A.M. Voorbij, G.H.A. Visser, H.W. Bruinse, and A. Franx, Risk factors for cardiovascular disease in women with a history of pregnancy complicated by preeclampsia or intrauterine growth restriction., *Hypertens. Pregnancy*. **26** (2007) 39–50. doi:10.1080/10641950601146574.

[5] T.L. Weissgerber, and L.M. Mudd, Preeclampsia and Diabetes, *Curr. Diab. Rep.* **15** (2015) 9. doi:10.1007/s11892-015-0579-4.

[6] R. McNally, A. Alqudah, D. Obradovic, and L. McClements, Elucidating the Pathogenesis of Pre-eclampsia Using In Vitro Models of Spiral Uterine Artery Remodelling, *Curr. Hypertens. Rep.* (2017). doi:10.1007/s11906-017-0786-2.

[7] J. Kieckbusch, L.M. Gaynor, and F. Colucci, Assessment of Maternal Vascular Remodeling During Pregnancy in the Mouse Uterus., *J. Vis. Exp.* (2015) 1–7. doi:10.3791/53534.

[8] P. Kaufmann, S. Black, and B. Huppertz, Endovascular Trophoblast Invasion: Implications for the Pathogenesis of Intrauterine Growth Retardation and Preeclampsia, *Biol. Reprod.* **69** (2003) 1–7. doi:10.1095/biolreprod.102.014977.

[9] T. Chaiworapongsa, P. Chaemsaithong, L. Yeo, and R. Romero, Pre-eclampsia part 1: current understanding of its pathophysiology, *Nat. Publ. Gr.* **10** (2014) 466–480. doi:10.1038/nrneph.2014.102.

[10] T. Clausen, S. Djurovic, F.R. Brosstad, K. Berg, and T. Henriksen, Altered circulating levels of adhesion molecules at 18 weeks' gestation among women with eventual preeclampsia: indicators of disturbed placentation in absence of evidence of endothelial dysfunction?, *Am. J. Obstet. Gynecol.* **182** (2000) 321–5.doi: 10.1016/S0002-9378(00)70218-3.

[11] P.K. Agatisa, R.B. Ness, J.M. Roberts, J.P. Costantino, L.H. Kuller, and M.K. McLaughlin, Impairment of endothelial function in women with a history of preeclampsia: an indicator of cardiovascular risk., *Am. J. Physiol. Heart Circ. Physiol.* **286** (2004) H1389-93. doi:10.1152/ajpheart.00298.2003.

[12] G. Valdés, Preeclampsia and cardiovascular disease: interconnected paths that enable detection of the subclinical stages of obstetric and cardiovascular diseases., *Integr. Blood Press. Control.* **10** (2017) 17–23. doi:10.2147/IBPC.S138383.

[13] L. Bellamy, J.-P. Casas, A.D. Hingorani, and D.J. Williams, Pre-eclampsia and risk of cardiovascular disease and cancer in later life: systematic review and meta-analysis, *BMJ.* **335** (2007) 974–974. doi:10.1136/bmj.39335.385301.BE.

[14] B. Arabin, and A.A. Baschat, Pregnancy: An Underutilized Window of Opportunity to Improve Long-term Maternal and Infant Health-An Appeal for Continuous Family Care and Interdisciplinary Communication., *Front. Pediatr.* **5** (2017) 69. doi:10.3389/fped.2017.00069.

[15] L.J. Alma, A. Bokslag, A.H.E.M. Maas, A. Franx, W.J. Paulus, and C.J.M. de Groot, Shared biomarkers between female diastolic heart failure and pre-eclampsia: a systematic review and meta-analysis, *ESC Hear. Fail.* **4** (2017) 88–98. doi:10.1002/ehf2.12129.

[16] E.K. Mallory, C. Zhang, C. Ré, and R.B. Altman, Large-scale extraction of gene interactions from full-text literature using DeepDive, *Bioinformatics*. **32** (2015) btv476. doi:10.1093/bioinformatics/btv476.

[17] I. Segura-Bedmar, P. Martínez, and C. de Pablo-Sánchez, A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents, *BMC Bioinformatics*. **12** (2011) S1. doi:10.1186/1471-2105-12-S2-S1.

[18] B. Percha, and R.B. Altman, A global network of biomedical relationships derived from text., *Bioinformatics*. **34** (2018) 2614–2624. doi:10.1093/bioinformatics/bty114.

[19] CH Wei, HY Kao, and Z Lu, Pubtator: a web-based text mining tools for assisting biocuration. *Nucleic Acid Res.* **41** (2013) (W1):W518-22. doi:10.1093/nar/gkt441

[20] DW Huang, BT Sherman, RA Lempicki. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc.* **4** (2009) 4-57.

[21] A Fabregat, K Sidiropoulos, P Garapati, M Gillespie, K Hausmann, R Haw, B Jassal, S Jupe, F Korninger, S McKay, L Matthews, B May, M Milacic, K Rothfels, V Shamovsky, M Webber, J Weiser, M Williams, G Wu, L Stein, H Hermjakob, P D'Eustachio. The Reactome pathway Knowledgebase. *Nucleic Acids Res.* **4** (2016) (D1):D481-7. doi: 10.1093/nar/gkv1351

[22] R. Satou, H. Penrose, and L.G. Navar, Inflammation as a Regulator of the Renin-Angiotensin System and Blood Pressure, *Curr. Hypertens. Rep.* **20** (2018) 100. doi:10.1007/s11906-018-0900-0.

[23] R. Carnagarin, V. Matthews, M.T.K. Zaldivia, K. Peter, and M.P. Schlaich, The bidirectional interaction

between the sympathetic nervous system and immune mechanisms in the pathogenesis of hypertension, *Br. J. Pharmacol.* (2018). doi:10.1111/bph.14481.

Address for correspondence

Dr Guillermo Lopez Campos.
Email: G.LopezCampos@qub.ac.uk