EREO: An Effective Rule Evaluation Framework for Discovering Interesting Patterns in US Birth Data and Beyond

Abhilash C. B.[1][a] and Kavi Mahesh[2][b]

[1]*Indian Institute of Information Technology Dharwad, IIIT Dharwad, India*
[2]*Department of Data Science and Intelligent Systems, Indian Institute of Information Technology Dharwad,*

Abstract: Birth data holds immense importance in healthcare for several reasons. It offers a comprehensive and representative sample of the population, enabling the identification of patterns and trends that can significantly impact public health policies and interventions. However, extracting interesting patterns from the vast birth data attributes poses a domain-specific and challenging problem. We can derive intriguing patterns by utilizing rare rules for identifying interesting associations. The level of interestingness depends on various factors, including the user, data, and domain. To address this, we propose the Effective Rule Evaluation using Ontology (EREO) framework, which incorporates two modes of rule evaluation. Firstly, the Integrated Rule Information Content (IRIC) measure is employed to quantify the level of interestingness. Secondly, the interesting rules are assessed by domain experts. The combined approach of these two modes of evaluation confirms the level of interestingness of the derived rules. The study demonstrates a significant relationship between these two modes of assessment, providing evidence of the convergence between expert evaluations and the ontology-based association rule measurements. This connection adds further value to the field by contributing to the understanding and measurement of interestingness within the context of ontology-based association rules

1 INTRODUCTION

Birth data is a valuable resource in healthcare research and analysis, providing a comprehensive and representative sample of the population (Abhilash and Mahesh, 2022). The patterns and trends identified in birth data can have significant implications for public health policies and interventions. However, extracting interesting patterns from the vast array of birth data attributes presents a challenging and domain-specific problem (C and Mahesh, 2021; Zhou et al., 2020; Tandan et al., 2021).

In this study, our focus is on addressing this challenge by leveraging rare rules to derive interesting associations and patterns from birth data (Abhilash and Mahesh, 2023). The concept of interestingness is subjective and depends on factors such as user preferences, data characteristics, and the specific domain of investigation. To address this, we propose an Effective Rule Evaluation using Ontology (EREO) frame-

work, which utilizes two modes of rule evaluation (Abhilash and Mahesh, 2023).

The first mode of evaluation involves the Integrated Rule Information Content (IRIC) measure (Manda et al., 2015), which quantifies the level of interestingness of the derived rules. By considering the information content embedded within the rules, IRIC helps identify the most intriguing and relevant patterns. The second mode of evaluation incorporates domain experts' assessment of interesting rules. Their expertise and insights play a vital role in confirming and validating the interestingness level of the derived patterns (Manda et al., 2013; Bringmann et al., 2011; Geng and Hamilton, 2006).

This study aims to demonstrate the efficacy and reliability of the EREO framework in evaluating the interestingness of rules derived from birth data. By employing both quantitative measures and expert evaluation, we aim to provide a systematic approach to identifying and interpreting interesting patterns.

Through extensive experimentation and evaluation, we establish a significant relationship between

[a] https://orcid.org/0000-0002-4864-8485
[b] https://orcid.org/0000-0002-4353-725X

the two rule assessment modes, confirming the proposed framework's effectiveness. The results of this study contribute to the field of healthcare analytics by enhancing our understanding of patterns and associations within birth data and providing valuable insights for public health decision-making.

Our contributions are as follows:

- We propose an Effective Rule Evaluation using Ontology (EREO) framework for evaluating the level of interestingness.

- We introduce an effective measure called Integrated Rule Information Content using Ontology (IRIC) for calculating the interestingness score.

Furthermore, the interesting rules are assessed by domain experts to ensure correctness, taking into consideration the parameter of clinical relevance.

Overall, this research highlights the importance of birth data analysis and proposes a novel framework, EREO, for evaluating the interestingness of association rules. By leveraging rare rules and incorporating expert evaluation, the framework enables the identification of meaningful patterns in birth data, offering potential benefits for public health policies, interventions, and improved healthcare outcomes.

The remainder of the paper is structured as follows: Section 2 discusses the recent literature. Section 3 presents the data and methodology for deriving interesting inferences from US birth data. Section 4 showcases the results along with the interestingness scores. Finally, Section 5 concludes the paper and highlights future research directions.

2 LITERATURE REVIEW

The literature surrounding the analysis of birth data and the evaluation of interesting patterns provides valuable insights into the existing research and methodologies in this field.

One area of focus in the literature is the exploration of association rule mining techniques for discovering patterns in birth data. Various studies have employed association rule mining algorithms to identify associations between maternal characteristics, birth outcomes, and demographic factors. These studies have shed light on risk factors for adverse birth outcomes, such as preterm birth and low birth weight, as well as the impact of socioeconomic factors on birth outcomes (Bekkar et al., 2020; Griggs et al., 2020; Robati et al., 2020).

Additionally, researchers have proposed different measures of interestingness to quantify the significance and relevance of discovered patterns. Measures like support, confidence, and lift have been commonly used to assess the interestingness of association rules in birth data. These measures aid in prioritizing and selecting the most meaningful and relevant patterns for further analysis and interpretation.

In recent years, the integration of ontologies in rule evaluation has gained attention in the literature. Ontologies provide a structured and standardized representation of domain knowledge, enabling the incorporation of domain expertise into the evaluation process. OntoVPA is a commercially available Dialogue Management System for Virtual Personal Assistants (VPAs) that utilizes ontologies and ontology-based rules (Wessel et al., 2019).

The approach for curating Gene Ontology (GO) data by utilizing association rule mining and introducing a specificity measure called VICD. The results demonstrate that VICD scores have stronger correlations with specificity values and lead to more consistent association rules (Shui and Cho, 2016).

(Mattiev and Kavsek, 2020), This study presents a new method for constructing accurate and compact classifiers by reducing the number of class association rules. The proposed associative classifier selects strong rules based on overall coverage and achieves high classification accuracy while generating smaller rules compared to traditional classifiers.

However, there remains a need for an effective framework that combines robust rule evaluation techniques, such as ontology-based measures, with expert evaluation to ensure the correctness and clinical relevance of the interesting rules. This gap in the literature highlights the importance of developing a comprehensive approach for evaluating the interestingness of association rules derived from birth data.

To address this gap, our study proposes the Effective Rule Evaluation using Ontology (EREO) framework, which leverages the Integrated Rule Information Content (IRIC) measure and domain expert evaluation. The IRIC measure accounts for the information content embedded within the rules, providing a quantitative indicator of interestingness. The involvement of domain experts ensures the assessment of interesting rules in terms of their clinical relevance and alignment with current healthcare practices.

By reviewing the existing literature, we have identified the need for an integrated approach that combines ontology-based measures, expert evaluation, and the domain-specific nature of interestingness assessment in the context of birth data analysis. Our study aims to contribute to this field by introducing the EREO framework and showcasing its effectiveness in evaluating the interestingness of association rules derived from US birth data.

3 DATA AND METHODS

This section provides a comprehensive overview of our research's data, ontology, and methodology.

3.1 Data and Ontology

In this study, we used the US birth data from 2020 as our primary dataset. The details of this dataset are presented in Table 1.

Table 1: Dataset and Ontology Information.

Dataset:	Birth Data 2020
Access Link:	[1]
Instances:	1 Million
Ontology:	BirthOnto
Link:	[2]

The birth data attributes encompass a wide range of information related to maternal characteristics, birth outcomes, demographic factors, and other relevant variables. These attributes include maternal age, race/ethnicity, education level, gestational age, birth weight, and geographical location. The inclusion of these attributes allows for a comprehensive analysis of various factors that may influence birth outcomes and patterns. To facilitate the evaluation of interestingness and enhance the interpretation of the derived association rules, we incorporated ontology into our methodology. Ontology provides a structured representation of domain knowledge and allows for the integration of expert-defined concepts and relationships. By incorporating ontology, we introduce a standardized framework that enables more meaningful and contextually relevant rule evaluation.

To illustrate the importance of ontology, consider an example where the birth data attributes include maternal education level and birth weight. By utilizing ontology, we can define and establish relationships between these attributes and relevant concepts, such as "high-risk pregnancy" or "low birth weight risk factors." This enables a more nuanced and detailed analysis of the association rules, leading to valuable insights into the relationships between maternal education level, birth weight, and their impact on birth outcomes.

3.2 Methodology

3.2.1 Hypothesis for Rare Rules

- Infants born to mothers who smoked cigarettes before pregnancy and had pre-pregnancy hypertension are more likely to be born preterm.
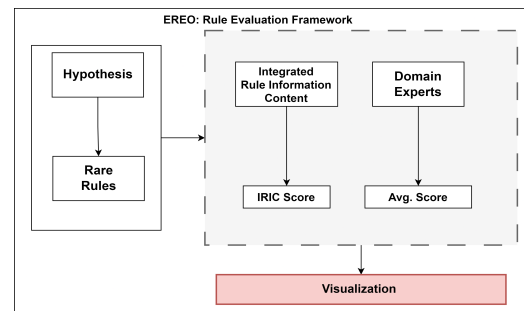

Figure 1: EREO Framework.

- Infants born to mothers aged more than 35 and had Hypertension Eclampsia have a higher likelihood of being born preterm.

- Infants born to mothers who had gonorrhoea and chlamydia during pregnancy have a higher likelihood of being admitted to intensive care.

- Infants born to mothers aged more than 35 who smoke have a higher likelihood of having Down syndrome.

- Mothers who had Failed External Cephalic Version and had Gestational Diabetes resulted in infants delivered through cesarean delivery.

- Mothers with a history of smoking and hypertension are at higher risk of developing cardiovascular diseases.

- Mothers who smoke during pregnancy are more likely to deliver infants with low birth weight than mothers who do not smoke during pregnancy.

3.2.2 Integrated Rule Information Content (IRIC)

IRIC (Integrated Rule Information Content) is a measure used to evaluate the interestingness of association rules in ontology-based data mining. It considers the information content of the predictor and outcome concepts in a rule, as well as the shared information between them. By calculating and combining the information content with weights, IRIC provides a comprehensive measure of the rule's interestingness. It is computed as shown in equation 1.

$$IRIC(X \rightarrow Y) = ((\alpha * N_{IC}(X)) + (\beta * N_{IC}(Y))) * N_{COMI}(X \rightarrow Y)$$
(1)

Where:

- $N_{IC}(X)$ is the information content of concept X
- $N_{IC}(Y)$ is the information content of concept Y
- $N_{COMI}(X \rightarrow Y)$ is the shared information between concepts X and Y
- α and β are weights assigned to concepts X and Y respectively, with $\alpha + \beta = 1$
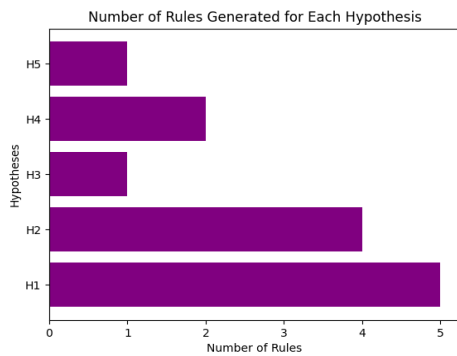
Number of Rules Generated for Each Hypothesis

Figure 2: Rules Generated for Hypothesis.

To calculate N_{IC}, you can use the formula:

$$N_{IC}(t) = -\log_2\left(\frac{p(t)}{UB(IC)}\right) \quad (2)$$

Where:

- t is a concept in the ontology
- $p(t)$ is the probability of t, calculated as the number of instances in the data that are annotated with t or any of its descendants, divided by the total number of instances
- $UB(IC)$ is the upper bound for information content, calculated as $-\log_2\left(\frac{1}{N}\right)$, where N is the total number of instances in the data

Once you have calculated N_{IC} and N_{COMI} for each rule, you can use *IRIC* to rank the rules in order of interestingness.

4 RESULTS AND DISCUSSION

The hypotheses defined in our study have been generated through collaborative consultations with doctors aimed at exploring specific relationships and patterns within the US Birth Data of 2020. These hypotheses provide valuable insights into potential associations between various factors and birth outcomes. The table below presents the hypotheses, along with their corresponding statuses. Table 2 represents the defined hypothesis status.

Figure 2 represent the number of interesting rules that were generated to support the hypothesis. The Accept hypotheses uncover significant correlations among various maternal characteristics, such as smoking, age, pre-pregnancy hypertension, gestational infections, and specific birth outcomes. These associations provide valuable insights into identifying risk factors and offer potential avenues for implementing interventions to address issues like preterm birth, intensive care admissions, and Down syndrome.

Table 2: Hypothesis from US Birth Data 2020.

#	Hypothesis	Status
H1	Infants born to mothers who smoked cigarettes before pregnancy and had pre-pregnancy hypertension are more likely to be born preterm.	Accept
H2	Infants born to mothers aged more than 35 and had Hypertension Eclampsia have a higher likelihood of being born preterm.	Accept
H3	Infants born to mothers who had gonorrhoea and chlamydia during pregnancy have a higher likelihood of being admitted to intensive care.	Accept
H4	Infants born to mothers over 35 who smoke have a higher likelihood of Down syndrome.	Accept
H5	Mothers who had Failed External Cephalic Version and had Gestational Diabetes resulted in infants delivered through cesarean delivery.	Accept
H6	Mothers with a history of smoking and hypertension are at higher risk of developing cardiovascular diseases.	Reject
H7	Mothers who smoke during pregnancy are more likely to deliver infants with low birth weight than mothers who do not smoke during pregnancy.	Reject

4.1 Semantic Interesting Rules

In the context of our work, the term "Semantic Interesting Rules" refers to association rules that incorporate the semantic knowledge captured by the ontology. In this study we have used Birthonto [3]. The results are represented in Table 3 4, and 5, the "Predictor_0" and "Predictor_1" columns indicate the antecedent part of the rule. The "Outcome" column specifies the consequent part of a rule. The "IRIC" column denotes the Integrated Rule Information Content, and "IS" column represents the Interestingness Score, indicating the degree of Interestingness of the rule. The "IRIC" values assigned to each rule in the table represent the Integrated Rule Information Content, which measures the level of interestingness of the rule. A higher IRIC value suggests a stronger as-

---

[3]https://bioportal.bioontology.org/

sociation between the predictor variables and the outcome. Additionally, the "'IS" column indicates the interestingness score of each rule, reflecting the degree of interest in the association. The IS is computed considering the IRIC values with the defined threshold.

The clinical significance of Table 3 resides in its depiction of the correlations between predictor variables and the occurrence of Neonatal Intensive Care Unit (NICU) admissions. The table specifically investigates the relationship between the predictor variables "CigarettesBeforePregnancyRecode" and "PrepregnancyHypertension" and their impact on NICU admission outcomes. Each row in the table represents a distinct scenario or range of "CigarettesBeforePregnancyRecode" values alongside their corresponding associations with NICU admissions, as outlined in the "Outcome" column.

Table 4 examines the relationships between maternal age and the presence of hypertension eclampsia in relation to the outcome of admission to the Neonatal Intensive Care Unit (NICU). It explores different age ranges (35-39 years, 50-54 years, 40-44 years, and 45-49 years) and their associations with NICU admission.

Table 5 represents the interesting rules generated based on the hypotheses H3, H4, and H5. The association between gonorrhoea and chlamydia with admission to the intensive care unit, smoking status and maternal age with confirmed Down syndrome, and failed external cephalic version and gestational diabetes with the cesarean delivery method are highlighted.

The distribution of IRIC values from the three tables reveals the relative interestingness of the rules. The rules with higher IRIC values are associated with lower distribution values, indicating higher interestingness.

Overall, utilizing the semantic information encapsulated within the ontology, we can uncover valuable insights and identify associations that might be overlooked by conventional rule-mining techniques. The semantic interesting rules provide a heightened understanding of the interrelationships between diverse attributes and outcomes within the birth data, enabling more informed decision-making in the realm of healthcare.

Figure 3 represents the boxplot that displays the distribution of the IRIC values derived from the data. By observing the boxplot, you can gain insights into the dispersion and distribution characteristics of the IRIC values.

Referring to Figure 4, the distribution of IRIC values from the three tables reveals the relative interest-
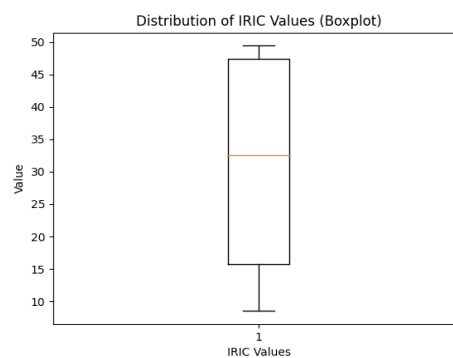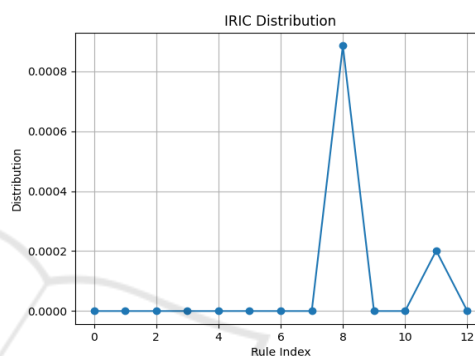

Figure 3: Box Plot of IRIC Values.

Figure 4: Distribution of IRIC Values.

ingness of the rules. The rules with higher IRIC values are associated with lower distribution values, indicating higher interestingness.

4.2 Domain Experts Evaluation

In our study, we recognized the importance of incorporating domain expert analysis to validate and evaluate the generated rules and their interestingness. We sought the expertise of doctors who served as domain experts in the field of healthcare and maternal care. Their insights and evaluations were invaluable in assessing the relevance and significance of the rules derived from the US Birth Data.

We employed an Interestingness Measurement Scale (IMS) proposed by (CB et al., 2023; Abhilash and Mahesh, 2023) to facilitate the evaluation process. The evaluation conducted by the domain experts revealed a significant positive relation between the Integrated Rule Information Content (IRIC) values and their expert evaluation. The evaluation was taken using google forms. Figure 5 represents the domain experts evaluation of all five hypothesis. Table 6 indicates the normalised IRIC scores with the average domain expert evaluation score of the interesting rules.

Table 3: Interesting Rule for Hypothesis-1.

Predictor_0	Predictor_1	Outcome	IRIC	IS
CigarettesBeforePregnancyRecode : 21-40	Pre-pregnancyHypertension : Yes	AdmissionToNICU : Yes	23.41	1
CigarettesBeforePregnancyRecode : 1-5	Pre-pregnancyHypertension : Yes	AdmissionToNICU : Yes	21.41	1
CigarettesBeforePregnancyRecode : 41 or more	Pre-pregnancyHypertension : Yes	AdmissionToNICU : Yes	24.26	1
CigarettesBeforePregnancyRecode : 6-10	Pre-pregnancyHypertension : Yes	AdmissionToNICU : Yes	19.15	1
CigarettesBeforePregnancyRecode : 11-20	Pre-pregnancyHypertension : Yes	AdmissionToNICU : Yes	18.33	1

Table 4: Interesting Rule for Hypothesis-2.

Predictor_0	Predictor_1	Outcome	IRIC	IS
MotherAge : 35-39 Years	HypertensionEclampsia : Yes	AdmissionToNICU : Yes	17.31	1
MotherAge : 50-54 Years	HypertensionEclampsia : Yes	AdmissionToNICU : Yes	38.39	1
MotherAge : 40-44 Years	HypertensionEclampsia : Yes	AdmissionToNICU : Yes	18.95	1
MotherAge : 45-49 Years	HypertensionEclampsia : Yes	AdmissionToNICU : Yes	7.03	0

Table 5: Interesting Rule for Hypothesis-3,4 and 5.

#	Predictor_0	Predictor_1	Outcome	IRIC	IS
H3	Gonorrhea : Yes	Chlamydia : Yes	AdmitToIntensiveCare : Yes	46.78	1
H4	SmokeStatus : Smoker	MotherAge : 40-44 Years	DownSyndrome : Confirmed	49.41	1
H4	SmokeStatus : Smoker	MotherAge : 35-39 Years	DownSyndrome : Confirmed	8.51	0
H5	FailedExternalCephalicVersion : Yes	GestationalDiabetes : Yes	FinalMethodOfDelivery : Cesarean	18.18	1

1. IRIC: Integrated Rule Infirmation Content.

2. IS: Interestingness Score.

3. Predictor: Antecedent of rule.

4. Outcome: Consequent of Rule.

Figure 6 indicates the comparison between the two levels of EREO framework. The RMS value of 0.534 suggests that there is some difference or discrepancy between the domain experts' scores and the normalized hypothesis scores. However, it is important to note that the magnitude of this difference is relatively small, indicating a moderate level of agreement between the two sets of scores. Also, as a statistical evaluation, we used T-test for the values in Table 6, it found that the p-value of 0.3103 indicates that there is a 31.03% probability of obtaining the observed difference in means by chance alone, assuming that there is no true difference between the two sets of scores.
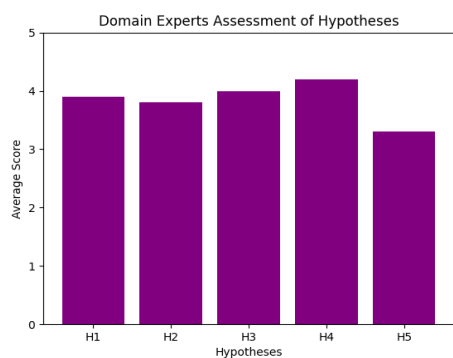
Figure 5: Domain Experts Evaluation of Rules.

Table 6: Comparison of Domain Experts Scores and Normalized Hypothesis Scores.

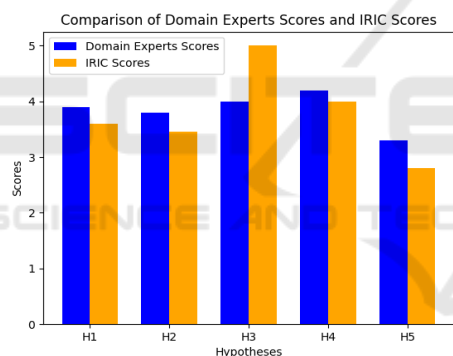| Hypotheses	Domain Experts Score	Normalized IRIC Score
H1	3.9	3.60
H2	3.8	3.45
H3	4.0	5.0
H4	4.2	4.03
H5	3.3	2.85

Figure 6: Comparision of IRIC and Domain Experts Evaluation.

4.3 Future Research Directions

Future research directions in this field involve the incorporation of additional measures for evaluating interestingness that align with the ontology-based association rule framework. By introducing AI-based evaluation as a third mode within the EREO framework, researchers can explore and leverage advanced techniques to further enhance the assessment of interestingness.

5 CONCLUSIONS

In summary, our study examined the relationships between various maternal characteristics and spe-

cific birth outcomes using US Birth Data from 2020. Through collaborative consultations with medical experts, we formulated hypotheses that provided valuable insights into these associations. By leveraging the semantic information embedded in the birth data ontology, we discovered deeper insights and uncovered connections that traditional rule mining techniques may overlook. The semantic interesting rules we generated enhanced our understanding of how different attributes relate to birth outcomes. These rules revealed important risk factors and potential interventions for preterm birth, intensive care admissions, Down syndrome, and other relevant outcomes. The clinical significance of these findings lies in their ability to guide healthcare decision-making and facilitate targeted interventions.

The insights gained from this study have implications for improving healthcare practices, identifying at-risk populations, and implementing preventive measures. Further exploration and analysis in this area can advance our knowledge and support evidence-based decision-making in the healthcare domain.

ACKNOWLEDGEMENTS

REFERENCES

Abhilash, C. and Mahesh, K. (2022). Ontology-based interestingness in covid-19 data. In *Metadata and Semantic Research: 15th International Conference, MTSR 2021, Virtual Event, November 29–December 3, 2021, Revised Selected Papers*, pages 322–335. Springer.

Abhilash, C. and Mahesh, K. (2023). Ontology is what makes data interesting: Interestingness framework for covid-19 corpora. *Journal of Information Science*.

Bekkar, B., Pacheco, S., Basu, R., and DeNicola, N. (2020). Association of air pollution and heat exposure with preterm birth, low birth weight, and stillbirth in the us: a systematic review. *JAMA network open*, 3(6):e208243–e208243.

Bringmann, B., Nijssen, S., and Zimmermann, A. (2011). Pattern-based classification: A unifying perspective. *arXiv preprint arXiv:1111.6191*.

C, A. and Mahesh, K. (2021). Graph analytics applied to covid19 karnataka state dataset. In *2021 The 4th Inter-*

national Conference on Information Science and Systems, page 74–80, New York, NY, USA. Association for Computing Machinery.

CB, A., Mahesh, K., and Sanda, N. (2023). Ontology-based semantic data interestingness using bert models. *Connection Science*, 35(1):2190499.

Geng, L. and Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR)*, 38(3):9–es.

Griggs, K. M., Hrelic, D. A., Williams, N., McEwen-Campbell, M., and Cypher, R. (2020). Preterm labor and birth: a clinical review. *MCN: The American Journal of Maternal/Child Nursing*, 45(6):328–337.

Manda, P., McCarthy, F., and Bridges, S. M. (2013). Interestingness measures and strategies for mining multi-ontology multi-level association rules from gene ontology annotations for the discovery of new go relationships. *Journal of biomedical informatics*, 46(5):849–856.

Manda, P., McCarthy, F., Nanduri, B., Wang, H., and Bridges, S. M. (2015). Information-theoretic interestingness measures for cross-ontology data mining. *arXiv preprint arXiv:1504.08027*.

Mattiev, J. and Kavsek, B. (2020). Coverage-based classification using association rule mining. *Applied Sciences*, 10(20):7013.

Robati, S., Saab, W., Durán-Retamal, M., Saab, W., Theodorou, E., Cawood, S., Serhal, P., and Seshadri, S. (2020). The association between elevated progesterone level on day of hcg trigger and live birth rates in art cycles: a single centre observational study. *Journal of Reproduction & Infertility*, 21(4):283.

Shui, Y. and Cho, Y.-R. (2016). Filtering association rules in gene ontology based on term specificity. In *2016 ieee international conference on bioinformatics and biomedicine (bibm)*, pages 1314–1321. IEEE.

Tandan, M., Acharya, Y., Pokharel, S., and Timilsina, M. (2021). Discovering symptom patterns of covid-19 patients using association rule mining. *Computers in biology and medicine*, 131:104249.

Wessel, M., Acharya, G., Carpenter, J., and Yin, M. (2019). Ontovpa—an ontology-based dialogue management system for virtual personal assistants. In *Advanced Social Interaction with Agents: 8th International Workshop on Spoken Dialog Systems*, pages 219–233. Springer.

Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., Gu, X., et al. (2020). Clinical course and risk factors for mortality of adult inpatients with covid-19 in wuhan, china: a retrospective cohort study. *The lancet*, 395(10229):1054–1062.