YSM28 2024 Abstracts

Abstracts are listed in alphabetical order by the last name of the presenter.

Abstract 1. Impact of COVID-19 Pandemic on Pathway Enrichment Analysis: Revelation of Biases and Implications for Disease Context

Name of presenter: Ahsani-Nasab, Sara; Italy

Abstract:

Pathway enrichment analysis tools are commonly employed to elucidate the molecular mechanisms and biological processes involved in various diseases. However, pathways databases and their semi-automatic methods linking genes to pathways could be prone to errors, especially during exceptional periods where a breakthrough discovery or a pandemic provokes a 'database shift' that can affect the entire literature. Given the unique combination of the Sars-CoV-2 pandemic and the subsequent fast-pacing publications of results, we investigated how these events affected the pathway enrichment analysis, aiming to uncover biases that might remain undetected during conventional periods.

We conducted a comprehensive analysis of COVID-related pathways using MeSH Disease terms from the full year 2018. PMIDs were retrieved and gene sets associated with each PMID were obtained. Multiple pathway analyses were performed using different databases. Significantly enriched COVID-associated pathways were identified based on adjusted p-values resulted from various p-value correction methods (<0.01). Comparative analyses within each database were conducted, considering adjusted p-values. We also evaluated the effectiveness of different p-value correction methods in the context of transcriptomics by comparing them within each database. All analyses were executed in R statistical language.

We demonstrated that for each database, over 20% of MeSH Disease terms were wrongly linked to COVID-related pathways. According to our findings, p-value correction methods resulted in different effects in different databases.

This study highlights a critical flaw in pathway enrichment analysis methods, revealing the potential for false discoveries and inaccurate associations between MeSH Disease terms and COVID-related pathways. Our findings challenge the common belief that there is one best p-value correction method for all conditions in transcriptomics e.g., all databases.

Abstract 2. Reviewing ITS Methods for the Exogenous Effect of COVID-19 on Infectious Disease Time Series

Name of presenter: Ali, Aqsa; Italy

Abstract:

On a global scale, the COVID-19 pandemic has caused significant disruptions, serving as an unprecedented exogenous event that has dramatically affected the time series data of several infectious diseases. These disruptions pose challenges to existing Interrupted Time Series (ITS) analysis methods, which are commonly used to examine the impact of external shocks on time dependent data. This study aims to review and evaluate the effectiveness of current ITS methodologies in capturing the immediate and long-term effects of the pandemic, as well as their ability to distinguish between pandemic-induced changes and other confounding factors. The insights from this review are intended

to guide the refinement and development of ITS techniques, ensuring more accurate and reliable interpretations of infectious disease trends in the context of COVID-19 and future global disruptions. The findings will be valuable for researchers and public health professionals seeking to mitigate the impact of large-scale disturbances on the dynamics of infectious diseases.

Keywords: Interrupted Time Series (ITS), COVID-19 Pandemic, Exogenous Effects, Infectious Disease Time Series, Pandemic Impact Analysis

Abstract 3. Uncovering the Constructs of Sentiment Analysis: A Validity Framework Approach

Name of presenter: Frantsits, Natalie; Austria

Abstract:

Sentiment analysis (SA) seeks to identify and quantify sentiments, moods, or feelings embedded within textual data. By 2024, SA models have become indispensable tools for uncovering social media trends and gauging public opinion on various topics, individuals, or events. Numerous (often premium) online platforms claim to provide key insights into brand perception, future trends, and market predictions based on sentiment measurements. However, these services, and the machine learning models they rely on, often lack transparency regarding the specific constructs they measure. Is it merely valence (i.e., happiness), the intensity of emotion, or an entirely different construct?

To address this ambiguity, the current study applies a construct validity framework, originally developed by Campbell and Fiske in the 1950s. This framework employs the multi-trait multi-method matrix - a restructured correlation matrix of measures derived from a common semantic network (i.e., network of meaning). The study examines sentiment measures from online SA models, alongside emotional measures grounded in established psychological theories (e.g., Ekman's basic emotions, Russell's Valence-Arousal model) and psychometrically validated scales for emotional experience. In addition to traditional linear models, dimensionality reduction techniques are utilized as an alternative approach to assessing construct validity.

If these SA models accurately represent sentiment (i.e., demonstrate construct validity), the analysis should reveal a single underlying factor, likely valence. However, drawing on psychological models of emotion, we anticipate the emergence of at least two dimensions of sentiment, with the possibility of a higher-dimensional or categorical model in light of recent developments in the field of emotion science.

This validity-based approach aims to elucidate the specific constructs captured by SA models, thereby contributing to much-needed transparency within the field. Moreover, by comparing the findings with real-world sentiment, this research lays the groundwork for the development of a future, validated "one model to rule them all."

Campbell, D. T. and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological bulletin, 56(2):81.

Abstract 4. Topic Modelling Classification of Basket and Umbrella Trials in pediatrics: an application of NLP-assisted Literature Review

Name of presenter: Khan, Mohd Rashid; Italy

Abstract:

Background: The development of personalized medicine has led to the emergence of innovative clinical trial designs such as basket and umbrella trials, which offer more flexible and efficient approaches for evaluating therapies in pediatric populations. Despite their potential, the literature on these trial designs, particularly in the pediatric context, is scattered and unstructured, making it challenging to

derive comprehensive insights. This study aimed to use Natural Language Processing (NLP) techniques, specifically topic modelling, to automatically extract and classify key topics from the literature on pediatric basket and umbrella trials.
Methods: A literature review of electronic databases including PubMed, Scopus, and Web of Science was conducted to gather relevant studies on pediatric basket and umbrella trials. After filtering and preprocessing the data, including the removal of stopwords, punctuation, and redundant terms, a Document-Term Matrix (DTM) was created from the cleaned text corpus. Latent Dirichlet Allocation (LDA) analysis was performed to identify the primary topics in the literature. LDA is an unsupervised machine-learning technique that automatically discovers latent topic structures within a collection of documents, allowing for the identification of key themes in the text. The optimal number of topics was determined using Deveaud's measure, and the results were validated against a manually classified subset of the literature.
Results: Topic modelling analysis revealed several dominant topics in the pediatric basket and umbrella trial literature, including issues related to trial design, endpoint selection, patient accrual, and the integration of Bayesian adaptive methods. The LDA model showed good agreement with the manually classified data, indicating the reliability of the automated approach for identifying relevant topics.
Conclusion: This study demonstrated the effectiveness of NLP techniques, particularly topic modelling, in extracting meaningful insights from large volumes of unstructured literature. By automating the classification and analysis of pediatric basket and umbrella trial publications, researchers can identify emerging trends and challenges in this critical area of clinical research more
efficiently. These findings can support more informed decision-making and the design of future pediatric trials.

Abstract 5. A new statistical index for evaluating variability in patient state index during pediatric anesthesia

Name of presenter: Khan, Noor Muhammad; Italy

Abstract:
Background: The Patient State Index (PSI) is a widely used tool for monitoring sedation levels in pediatric anesthesia, providing an indication of the patient's level of consciousness and depth of anesthesia. The stability of sedation during anesthesia is equivalent in statistical jargon to the stationarity of the process. However, the detection of time points when the PSI level changes from a stable sedation level remains a challenge. Since the distribution of PSI is usually non normal and is expected to have outliers, a robust method is required to evaluate the phases. The anesthesia period is clinically divided into well-defined five phases; however, recent studies have unexpectedly detected large variations of PSI have even within same phase during pediatric anesthesia.
Methods: This study proposes the Variability Ratio Index (VARI), a simple statistical tool based on the deviation of PSI from its stationary process, to evaluate sedation phases. We considered the probability of a change point, where PSI deviates, can vary across different phases. We used quasi-binomial distribution that describes additional variation of PSI from its stationary process, to develop VARI. Change points are detected using the pruned exact linear time algorithm. We also checked the robustness behavior of VARI in both parametric bootstrapping in Bayesian paradigm and Monte Carlo simulation.
Results: To demonstrate the practical application of VARI, a single-center retrospective study was conducted using PSI data from pediatric patients undergoing cardiac surgery with extracorporeal circulation. The study included twenty patients monitored using the Sedline monitor at 124,699 time

points. We observed large variation of PSI within each phase and VARI evaluated all phases with satisfactory results. VARI successfully identified the hypothermic phase with the lowest value and the awakening phase with the highest value of it, highlighting its potential in assessing sedation depth during anesthesia. VARI showed robust behavior in both parametric bootstrapping under Bayesian paradigm and Monte Carlo simulation. VARI converged into Beta distribution with parameters Shape 1 < Shape 2 in 10,000 iterative schemes. Finally, we applied Generalized Estimating Equation to capture the correlation structure of each patient to make patient wise predictions on the deviations of PSI.
Conclusions: We developed R package "varifinder" to facilitate the use of VARI. Further research and data are necessary to fully explore the utility of VARI in different clinical settings.
Keywords: VARI, PSI, robustness, GEE.

Abstract 6. Sharp oracle inequalities and universality of the AIC and FPE
Name of presenter: Köstenberger, Georg; Austria
Abstract:
In two landmark papers, Akaike introduced the AIC and FPE, demonstrating their significant usefulness for prediction. In subsequent seminal works, Shibata developed a notion of asymptotic efficiency and showed that both AIC and FPE are optimal, setting the stage for decades-long developments and research in this area and beyond. Conceptually, the theory of efficiency is universal in the sense that it (formally) only relies on second-order properties of the underlying process, but, so far, almost all (efficiency) results require the much stronger assumption of a linear process with independent innovations.
In this work, we establish sharp oracle inequalities subject only to a very general notion of weak dependence, establishing a universal property of the AIC and FPE. A direct corollary of our inequalities is asymptotic efficiency of these criteria.
Our framework contains many prominent dynamical systems such as random walks on the regular group, functionals of iterated random systems, functionals of (augmented) Garch models of any order, functionals of (Banach space valued) linear processes, possibly infinite memory Markov chains, dynamical systems arising from SDEs, and many more.

Abstract 7. Branching processes in nearly degenerate varying environment
Name of presenter: Kubatovics, Kata; Hungary
Abstract:
We investigate branching processes in varying environment, for which $\overline{f_n} \to 1$ and $\sum_{n=1}^{\infty}\left(1-\overline{f_n}\right)_+ = \infty$, $\sum_{n=1}^{\infty}\left(\overline{f_n}-1\right)_+ < \infty$, where $\overline{f_n}$ stands for the offspring mean in generation n.
Since subcritical regimes dominate, such processes die out almost surely, therefore to obtain a nontrivial limit we consider two scenarios: conditioning on non-extinction, and adding immigration. In both cases we show that the process converges in distribution without normalization to a nondegenerate compound-Poisson limit law.
We also prove functional limit theorems in the above cases. In the conditional setup, the limiting process is a simple birth and death process conditioned on non-extinction at time 1, while the process with immigration tends to a continuous time branching process with immigration with a simple birth and death branching counterpart.
The proofs rely on the shape function technique, worked out by Kersting (2020).

Abstract 8. Detection of Differential Item Functioning

Name of presenter: Lajos, Hanka; Hungary

Abstract:

Introduction

In the case of differential item functioning (DIF), a particular item of an ability assessment shows different characteristics in two groups of students. Students with the same level of ability have differing probabilities of correctly responding to the given item depending on their group membership. The focus of this study are various methods for detecting DIF.

Methods

DIF analysis was conducted on the set of common items for two consecutive mathematics assessments in the Hungarian assessments of competencies (OKM). Two indices of DIF (NCDIF and UAM) were used combined with multiple thresholds for diagnosing DIF and methods for linking the scales of the assessments. The aim was the detection of DIF items and the investigation of the effect of the chosen index, critical threshold and linking method on the diagnosis of DIF and the consequent linking function.

Outcome

The results of the multiple DIF analyses varied substantially in the number and identity of diagnosed DIF items. The two indices were able to detect different types of discrepancies between the item characteristic functions derived from the response patterns of each year's student population. However, the detected item set was largely insensitive to the applied linking method.

Conclusions

Exclusion of detected DIF items had a considerable impact on the final linking function calculated for linking the ability and parameter scales of the two assessments. This finding affirms an important application of DIF detection. This effect was mainly observable using one of the two indices (NCDIF). Omitting items diagnosed as differentially functioning according to the other index had only minor effect on the linking functions. The indices diagnosed largely different sets of items as DIF. However, the selection of the linking method had no effect on the detected item set in almost all combinations of chosen index and critical threshold.

Abstract 9. Evaluating the Agreement between Human Preferences, GPT-4V and Gemini Pro Vision Assessments

Name of presenter: Maltar, Jurica; Croatia

Abstract:

Since the emergence of large language models, we have been witnessing how artificial intelligence has reached the level of human understanding. In this presentation, we will provide insights into the analysis of the degree of agreement between human and large language model ratings of restaurants. We were curious whether a visual large language model, i.e. an LLM with the addition of visual input, can rate images of restaurants the way people do.

Methodologically, we conducted a survey providing respondents with interior and food images of nine restaurants and asked them questions about these restaurants. The respondents answered each question with a Likert-scale rate from 1 to 5. Then we provided the context of respondents, considering the confidentiality of the data, to GPT4-Vision and Gemini Pro LLMs and asked them the same questions with the same images. Having ratings from survey respondents, GPT4-Vision and Gemini Pro, we were able to compare the degree of agreement between human and LLM ratings by using standard descriptive statistics and the interclass correlation measure.

The outcome of the research is the comparison between answers from human respondents and LLMs. In general, GPT4-Vision shows a better correlation to human input compared to Gemini Pro. Also, answers that correspond to questions related to images of interior are more correlated than answers that correspond to questions related to images of food.
In conclusion, we will present an objective method for assessing agreement between LLMs and human ratings and the way we could prompt LLMs in order to achieve this objective, which could be useful in planning future studies.

Abstract 10. Similarity coefficients for two sets of binary classifiers: methodology and applications
Name of presenter: Perišić, Ana; Croatia
Abstract:
The presence of a certain trait is commonly assessed by different heuristic rules or by different binary classifiers. Binary classification is a common task and numerous methods have been developed for binary classification problems. Applying different binary classification rules or methods often results in different classifications. When we have two sets of binary classifiers, we observe two sets of binary vectors, and our goal is to assess the similarity of two such sets of vectors. Although measuring the similarity of two binary vectors is an active research field, measuring the similarity of two sets of binary vectors still offers limited methodological solutions. We present an approach for calculating the similarity within such a set of vectors and between two such sets of vectors, where we evaluate the similarity in consensus agreement. We propose a generalization of the k-adic definition of similarity and calculate the within-set similarity by applying the k-adic version of the Jaccard coefficient and the simple matching coefficient and extend these coefficients to compare two sets of binary vectors. We derive the large sample variances of the coefficients and present some desirable properties of the coefficients. We show several applications of the established coefficients and present two applications in more detail. The first application deals with conceptual uncertainties that arise from a lack of clarity in certain concepts, such as churn in the noncontractual business domain. We propose a methodology for evaluating the similarity of churn definitions and churn definition groups. The second application presents the methodology for calculating the similarity of two binary classification algorithms. For a given dataset, algorithms are evaluated on the basis of the predicted classes, where for each algorithm, we take into account different values of hyperparameters.

Abstract 11. Modelling the causal relationships between environmental factors and thyroid function using genome-wide association studies and Mendelian randomisation
Name of presenter: Pleić, Nikolina; Croatia
Abstract:
Introduction
In numerous observational studies, thyroid function has been linked to metabolic syndrome (MetS), as well as to serum 25-hydroxyvitamin D [25(OH)D] levels, but the direction of effects and the exact causal mechanisms remain unclear. We employed Mendelian randomization (MR) to examine the causal relationship between thyroid function and MetS, as well as to investigate the causal effect of serum 25(OH)D concentration on thyroid function indicators.
Methods
In our first study, we performed a two-sample MR analysis using summary statistics from the most comprehensive genome-wide association studies (GWAS) of thyroid-stimulating hormone (TSH), free thyroxine (fT4), MetS and its components. In our second study, in addition to TSH and fT4, we included

free triiodothyronine (fT3), total triiodothyronine (TT3), thyroid peroxidase antibody levels (TPOAb), low TSH, high TSH, autoimmune hypothyroidism and hyperthyroidism. We used the multiplicative random-effects inverse variance weighted (IVW) method for primary analysis, supplemented by weighted mode, weighted median, MR-Egger, MR-PRESSO, and CAUSE methods.

Results

Our results indicate that higher fT4 levels are causally linked to a lower risk of developing MetS (OR = 0.96, p = 0.037). Genetically predicted fT4 was also positively associated with HDL-C (β = 0.02, p = 0.008), while genetically predicted TSH was positively associated with TG (β = 0.01, p = 0.044). Reverse MR analysis showed that genetically predicted HDL-C was negatively associated with TSH (β = −0.03, p = 0.046). IVW, MR Egger, and CAUSE analyses suggested a causal effect of 25(OH)D on high TSH, with each 1 SD increase in serum 25(OH)D linked to a 12% decrease in risk of high TSH (p = 0.02). Additionally, MR Egger and CAUSE analyses suggested that a 1 SD increase in serum 25(OH)D was associated with a 16.34% decrease in risk of autoimmune hypothyroidism (p = 0.02).

Conclusion

Our study suggests that variations in normal-range thyroid function are causally associated with the diagnosis of MetS and with lipid profile. Our results further indicate that higher genetically predicted vitamin D levels may causally reduce the likelihood of having high TSH or autoimmune hypothyroidism. However, no causal influence of vitamin D on other thyroid parameters was observed.

Abstract 12. Evaluating Large Language Models as Survey Respondents

Name of presenter: Rakovics, Zsófia; Hungary

Abstract:

The emergence of large language models (LLMs) has created a new opportunity for social scientific research methods. For both qualitative and quantitative empirical research where language mediates information gathered from people, it has become a realistic possibility to generate data using virtual respondents simulated by LLMs.

The quality of the data generated by LLMs depends largely on the way it is extracted, thus the methodology of prompt engineering – finding the best inputs for the desired outputs – has been rapidly developing. A critical question in extracting data for social research purposes is which prompt should be used to define the context which activates the appropriate patterns in the model to get relevant responses. The potential impact of LLM-generated data necessitates its critical methodological analysis. From a positivist perspective, if the methodology of virtual data collection can be developed, it could aid in preparing for human data collection with virtual pilot studies and support a wider scope for improving and supplementing (e.g. by imputation) the real data collected.

Using LLMs (GPT-3.5-turbo, GPT-4-turbo, Llama-2-70b, Mixtral-8x7B), we generated answers for politics and democracy related attitude questions of the European Social Survey (10^{th} wave) and statistically compared these to the real responses. We explored different prompting techniques (e.g. zero-shot, few-shot) and the effect of different types and richness of contextual information provided to the models. The results suggest that the tested LLMs generate realistic answers and are good at invoking the expected patterns from limited contextual information, but struggle in a zero-shot setting. A critical perspective is essential to ensure that known biases of LLMs do not remain unexplored in these applications, and even more so to consciously consider the social reality that is not represented in the linguistic space of the Internet.

Abstract 13. Improving AI explainability through experimental design
Name of presenter: Stadler, Alexandra; Austria
Abstract:
Local Interpretable Model-agnostic Explanation (LIME) is a popular technique in explainable AI designed to provide interpretability to complex machine learning models. LIME is an algorithm that perturbs the input data and fits a local model to the new artificial sample in order to generate an explainable model in the neighbourhood of an instance of interest. An explanation can be any model that is simple enough for a human to interpret, for instance, we concentrate on local linear regression models. While LIME is widely used, there are a number of pitfalls in this method, among them is the randomized sampling of new input data.
We aim to improve LIME's sampling scheme on tabular data by employing design of experiments for locally weighted linear regression. We compute D- and A-optimal designs for local linear models given an instance of interest and use the resulting designs to generate new efficient samples to fit a local model. This serves to reduce the number of new samples that have to be generated, while minimizing the (generalized) variance of the estimators in the local linear model.
In an application where the machine learning model's predictions are costly, this reduces the computational effort and saves time. Additionally, this approach leads to deterministic explanations, whereas in LIME, randomized sampling leads to variability in explanations.
We conclude that LIME's standard method of perturbing an instance of interest is subject to randomness, which can lead to instability of the resulting explanations. Further, this sampling method is costly in terms of the machine learning model's number of predictions that have to be generated. An efficient approach using design of experiments can help to overcome these difficulties.

Abstract 14. Leveraging Cancer Incidence for Lead Time Estimation in Cancer Screening Programmes
Name of presenter: Vratanar, Bor; Slovenia
Abstract:
In cancer screening programmes, participants are regularly screened every few years using blood tests, urine tests, or medical imaging to detect cancer at an earlier time, when it is presumed to be more curable. Without screening, cancer would likely progress undetected until symptoms appear. The interval between early detection and the eventual onset of symptoms, had screening not been conducted, is known as lead time. Understanding lead time is essential for better planning and timing of treatment interventions.
Estimating lead time is challenging because it is a hypothetical random variable that can only be inferred indirectly. In our study, we introduce a novel method for estimating lead time, using a data source previously untapped for this purpose—cancer incidence. We hypothesize that earlier detection of cancer due to screening should result in observable shift in cancer incidence rates, stratified by age and year of diagnosis. Our method leverages this information, and estimates lead time using a maximum likelihood estimator. In principle, the user specifies the distribution of lead time, and the method finds the parameters that best fit the observed shift in cancer incidence.
Our approach is flexible, allowing for the inclusion of additional covariates and accounting for overdiagnosis. The data required for this method are routinely available from cancer registries and provided by population tables, making it easy for implementation. We validated our method through simulations and applied it to data from the Slovenian breast cancer screening programme, demonstrating its effectiveness and utility.