Empowering Colorectal Cancer Research through Advanced Data Integration and Analysis: A case study of the DIOPTRA project

Marilena Tarousi[1[0000-0003-0835-3751]], Stavros-Theofanis Miloulis[1[0000-0001-5091-5530]], Maria Haritou[1[0000-0003-1136-8209]], Konstantinos Bromis[1[0000-0002-6176-2282]], Ioannis Kouris[1[0000-0003-3848-3979]], George Botis[1[0009-0008-2950-0929]], Ioannis Kakkos[1[0000-0001-8365-2140]], and George Matsopoulos[1[0000-0002-2600-9914]]

[1] Biomedical Engineering Laboratory, School of Electrical and Computer Engineering, National Technical University of Athens, 15780 Athens, Greece
mtarousi@biomed.ntua.gr, smiloulis@biomed.ntua.gr, mhari@biomed.ntua.gr, konbromis@biomed.ntua.gr, ikouris@biomed.ntua.gr, botis_g@biomed.ntua.gr, ikakkos@biomed.ntua.gr, gmatsopoulos@biomed.ntua.gr

Abstract. The escalating incidence of colorectal cancer (CRC), paired with the progressively decreasing age threshold and the low screening adherence by citizens, underscores the urgent need for innovative approaches to screening and early detection. In this light, the DIOPTRA EU project presents a pioneering initiative aimed at harnessing advanced data collection, integration, and analytics methodologies to unlock critical insights into CRC. This paper provides a comprehensive overview of the DIOPTRA platform's conceptual architecture and its potential implications for addressing CRC research challenges. Leveraging data integration, artificial intelligence, and state-of-the-art biomarker analysis, DIOPTRA offers a promising solution for enhancing early detection and thus improving patient outcomes. The integration of the DIOPTRA Back-end, facilitating data ingestion, curation, and storage, with the Front-end's user-friendly interfaces and a dedicated anonymization tool, offers the required link between clinical practice and innovative research in CRC screening programs. By fostering interdisciplinary partnerships and embracing continuous innovation, DIOPTRA has the potential to revolutionize CRC screening practices, reduce mortality rates, and shape the future of healthcare delivery. In this regard, this paper stresses the importance of ongoing efforts to advance cancer screening technologies, highlighting the role of integrated platforms like DIOPTRA in improving public health outcomes through multidisciplinary research within clinical settings.

Keywords: Colorectal Cancer Screening, Early Detection, Preventive Healthcare, Artificial Intelligence, Data Integration, Healthcare Informatics.

1 Introduction

Excellence in cancer prevention and early detection demands continuous innovation, particularly in addressing the formidable challenge posed by colorectal cancer (CRC). With its significant global burden, CRC stands as the third most common tumor in men and the second in women, accounting for a substantial portion of cancer-related morbidity and mortality worldwide [1, 2]. The escalating incidence of CRC, projected to reach alarming levels by 2040, underscores the urgency for proactive strategies. Current statistics reveal a distressing reality, with 1.9 million new cases diagnosed in 2020 alone, contributing to an estimated 0.9 million deaths globally [2]. Moreover, CRC's insidious nature is exemplified by its propensity for metastasis, with a substantial proportion of cases presenting with metastatic disease at the time of diagnosis or experiencing metastatic relapse during the disease course [3].

Amidst the formidable challenge posed by CRC, the current arsenal of treatment modalities spans a spectrum of interventions encompassing surgical excision, chemotherapy, targeted therapies, and immunomodulatory agents [2]. However, the cornerstone of CRC control lies in preventive strategies, particularly population-based screening initiatives. Endoscopic (e.g. colonoscopy) and non-invasive (e.g. fecal immunochemical test - FIT) screening modalities have demonstrated efficacy in detecting early-stage disease, leading to a substantial reduction in mortality rates [4, 5].

Despite the strides made in CRC prevention and management, significant challenges persist, hindering the realization of optimal outcomes. Factors such as a) modern lifestyle favoring CRC incidence [6], b) suboptimal screening participation (mainly attributed to limited accessibility to screening services [7] and the taxing nature of existing screening modalities [8]), and c) low sensitivity of current methods for pre-malignancies [9], contribute to the ongoing burden of CRC-related morbidity and mortality. Furthermore, disparities in risk factor assessment [10] and the underutilization of emerging technologies, such as liquid biopsy [11], foreground the need for comprehensive, integrated approaches to CRC screening in an era where the rising incidence of early-onset CRC is rendering the current age thresholds outdated [12].

In response, the DIOPTRA project (Early Dynamic Screening for Colorectal Cancer via Novel Protein Biomarkers Reflecting Biological Initiation Mechanisms) emerges as a pioneering initiative aimed at harnessing advanced data collection, integration, and analytics to unlock critical insights into CRC. By leveraging multidisciplinary expertise in clinical practice, proteomics, and informatics, DIOPTRA seeks to revolutionize our comprehension of CRC and enhance early detection as well as prevention. Within this scope, effective data collection methodologies are pivotal in comprehending complex diseases such as CRC, as they establish the foundation for subsequent analyses, facilitating the identification of key risk factors, elucidation of disease mechanisms, and development of preventive strategies. This paper contributes to this pursuit by providing an overview of the data collection methodology employed within DIOPTRA, paired with an integrated architecture that enables homogenized data analysis. It delineates the systematic approach to data structuring, anonymization, upload and assessment processes, as well as software development and integration.

2 DIOPTRA Vision and Innovation

DIOPTRA embodies a visionary approach to revolutionizing colorectal cancer screening and prevention in everyday medical practice, driven by a commitment to accessibility, innovation, and tangible healthcare impact. It envisions to serve as a front-line clinical decision support tool, leveraging both risk factors and novel protein biomarkers to identify high-risk cases requiring colonoscopy assessment. This approach fosters targeted allocation of resources, eliminating additional burden due to unnecessary invasive procedures and overcoming barriers of traditional screening methods [9]. Overall, the project vision encompasses the following key elements:

Minimally-Invasive Liquid Biopsy. Against the backdrop of taxing procedures hindering adherence to scheduled screening, DIOPTRA envisions a straightforward blood test for stratification leveraging novel protein biomarkers. By seamlessly integrating CRC screening into existing minimally invasive procedures such as standard bloodwork, the evaluated population can be significantly broadened, thereby increasing participation rates and bypassing age screening thresholds.

Risk-Factor-Based Assessment. DIOPTRA integrates a holistic risk factor model, harnessing cutting-edge analytics towards efficient risk assessment. This will, by extension, enable the provision of personalized behavioral recommendations aiming to mitigate pre-malignant stages and reduce progression risk.

Biologically Relevant Evidence and Explainability. DIOPTRA recognizes the importance of medical explainability and transparency, particularly in the context of decision support. Hence, our approach emphasizes the incorporation of biologically relevant evidence alongside AI insights, ensuring the reliability and interpretability of diagnostic recommendations. In this spirit, to validate DIOPTRA markers as key regulatory nodes in CRC development, protein networks will be constructed, aiming to elucidate the underlying molecular mechanisms of CRC pathogenesis.

Innovation lies at the heart of DIOPTRA's mission to redefine clinical standards in CRC prevention and early detection, making screening faster, more precise, personalized, accessible, and affordable for a broader population. This approach not only enhances screening accessibility, but also prioritizes safety by minimizing contact between individuals and healthcare professionals, particularly crucial in times of pandemic outbreaks like the COVID-19 crisis. Moreover, DIOPTRA's cost-effective protocol ensures long-term sustainability, contrasting with resource-intensive methods associated with high PPPY costs (per patient per year) [8]. Inspired by successful precedents such as the FDA-approved Epi proColon® test [11] which relies on a single validated biomarker, DIOPTRA sets a benchmark for innovation in CRC risk assessment, promising earlier detection and prevention strategies.

Beyond simply improving screening methods, DIOPTRA's innovation extends to reshaping research and knowledge on cancer. By undertaking comparative validation of biomarkers using paired blood & tissue samples from different population groups [13] (healthy, non-advanced adenomas, advanced adenomas, CRC patients), DIOPTRA bridges a crucial gap in the field, enabling reliable assessment of protein sensitivity for CRC screening. Moreover, the combined implementation of Next Generation Sequencing (NGS) and AI techniques facilitates a deeper understanding of CRC mechanisms, enabling us to elucidate the regulatory contributions of markers in protein network pathways. This holistic approach bears the potential not only to enhance diagnostic accuracy, but also to provide valuable insights into the underlying development of CRC, irrespective of screening sensitivity. Through validated AI systems and the expansion of risk factor evaluation to encompass lifestyle and regional aspects, DIOPTRA aims to set a new standard for CRC screening, promising a paradigm shift in cancer early detection and prevention strategies.

3 Data Management and Analysis in DIOPTRA

Data anonymization and management are paramount within the DIOPTRA project to safeguard the privacy, integrity, and seamless homogenization of sensitive clinical data. Employing a dedicated anonymization tool, DIOPTRA utilizes the k-anonymity method with the Mondrian algorithm [14] to obscure individual identities within datasets while preserving data utility. This advanced anonymization approach ensures compliance with stringent privacy regulations and ethical considerations governing the study, while also reinforcing the clinicians' trust to the study process and subsequent data analysis.

The structured data upload process, guided by a project-specific data ontology, robust algorithms and stringent validation checks, guarantees the accuracy, completeness, and adherence to project criteria of uploaded datasets. Specifically, the DIOPTRA data template is used to create structured datafiles within the clinical sites, which are then fed to the anonymization tool before being uploaded to the DIOPTRA Back-end using the Front-end functionalities (**Fig. 1**). In this regard, the DIOPTRA Back-end component also plays a pivotal role in ensuring data integrity and consistency through comprehensive curation processes before data storage, rendering the data fit for analysis and interpretation. Overall, the DIOPTRA Software provides modules for data ingestion, error detection, transformation, curation, and cataloguing operations, offering clinicians user-friendly interfaces for uploading electronic health records' (EHR) data and reviewing ingested data from clinical sites. Through these meticulous data management practices, the DIOPTRA project establishes a secure, efficient, and standardized framework for handling clinical data, fostering compatibility and interoperability across different clinical sites, thereby facilitating meaningful insights and analyses in the quest for improved CRC screening methodologies.

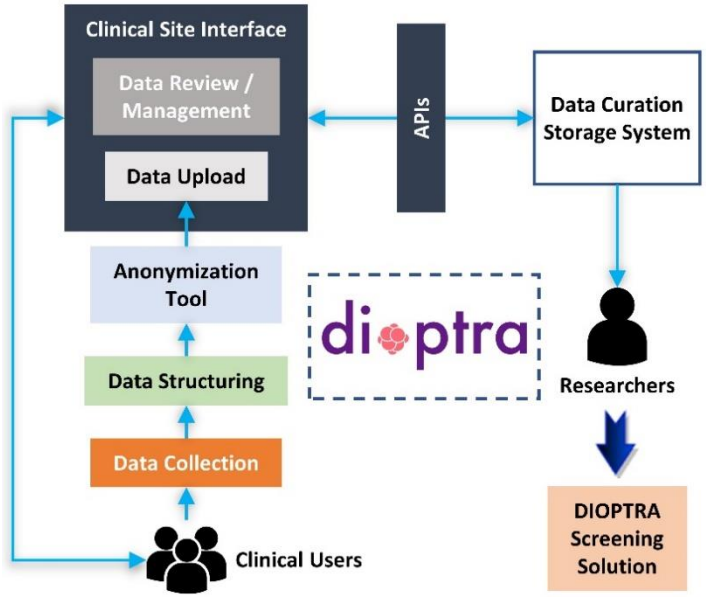

Fig. 1. DIOPTRA Infrastructure & User Interaction

Within the project architecture facilitated by the DIOPTRA Software, Artificial Intelligence (AI) stands as a pivotal component, revolutionizing CRC screening methodologies. Leveraging AI techniques to reinforce blood biomarker analysis and risk assessment within clinical decision support, DIOPTRA enhances detection accuracy and risk prediction in CRC screening. Particularly, state-of-the-art AI will be applied on the protein features derived via bioinformatics analysis, as well as on the risk factor data – both independently and in combination – to a) identify the biomarker subset with optimal diagnostic capacity, b) develop an AI-assisted non-invasive risk assessment model, and c) propose a holistic screening tool incorporating both risk factor data and blood-derived results. In this process, the use of Explainable Artificial Intelligence (XAI) through Knowledge-Graph-based (KG) modelling [15] will foster trust and acceptance among stakeholders.

4 Architectural Design of the DIOPTRA Platform

4.1 DIOPTRA's Conceptual Architecture

DIOPTRA's methodology is structured into distinct phases (as illustrated in **Fig. 2**), each building upon the outputs of the preceding phase to advance towards its ultimate goal. Phase 1 initiates with the preliminary data collection, encompassing paired tissue and liquid biopsy samples along with aggregated demographic, behavioral, environmental, medical, and family history information based on the DIOPTRA data model. This data is standardized and organized into a homogenized dataset, serving as the foundation for subsequent phases.

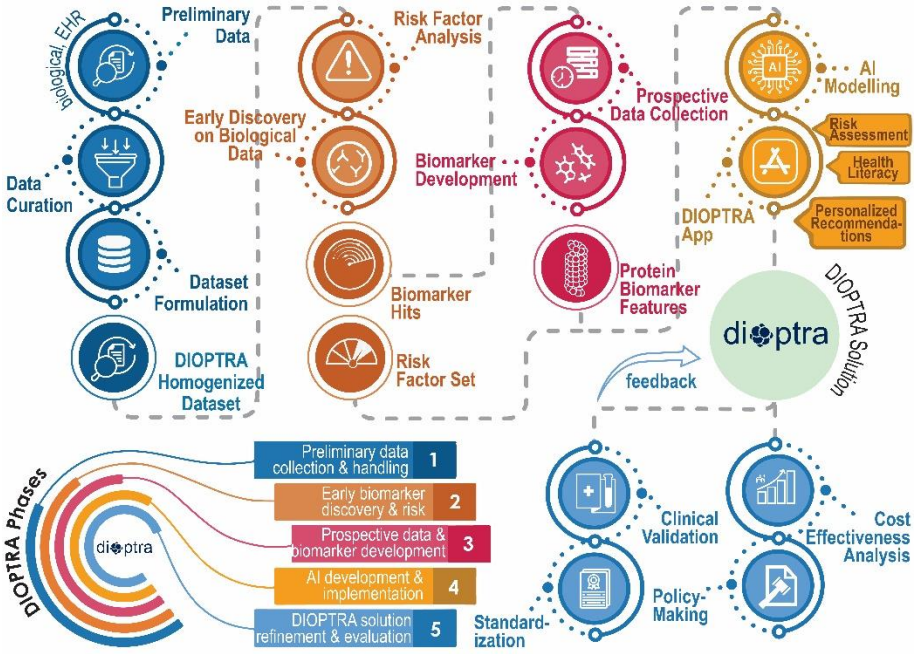

Fig. 2. DIOPTRA Methodological Architecture ones are justified.

In Phase 2, the focus shifts to investigating risk factors and conducting an initial analysis of paired samples for early biomarker discovery. This phase aims to establish a holistic set of CRC risk factors and identify potential biomarker candidates for further evaluation. Phase 3 transitions into prospective data collection, expanding the dataset with additional risk factor information and biological samples collected from citizens visiting DIOPTRA's clinical sites for screening services, based on a jointly approved clinical protocol. Phase 4 focuses on AI-based analysis, utilizing risk factors and protein features to develop comprehensive stratification models. Additionally, knowledge gleaned from this analysis informs the development of the DIOPTRA mobile app, enabling personalized behavioral recommendations. In Phase 5, the full DIOPTRA solution undergoes clinical validation, real-settings' feedback, and refinement within the prospective study. This stage seeks to validate the efficacy of the DIOPTRA solution in real-world healthcare settings and identify implementation needs for scaling up its adoption. Finally, cost-effectiveness analysis against the current screening practices will be conducted to provide quantified evidence for potential policy changes in CRC screening guidelines. By following this systematic approach, DIOPTRA aims to deliver a comprehensive and effective solution for CRC screening and early detection, ultimately contributing to improved patient outcomes and healthcare policies.

4.2 Integration & Testing of DIOPTRA Platform

The integration and testing phase of the DIOPTRA platform are pivotal stages in ensuring its efficacy, reliability, and adherence to the project objectives. This phase involves a systematic approach to seamlessly amalgamate diverse components, validate system performance, and refine functionalities to meet project requirements. During this phase, the DIOPTRA Back-end, responsible for data ingestion, retrieval, curation, and storage, is meticulously integrated with other platform modules. Similarly, the DIOPTRA Front-end, featuring the clinical site interface operating in compatibility with the data template and the anonymization tool, is seamlessly incorporated with the Back-end to facilitate smooth data upload and management processes.

Moreover, integral to the integration process is the establishment of API gateways, granting seamless communication between the different platform components. These gateways undergo rigorous testing to ensure their reliability, consistency, and interoperability, enabling smooth data exchange and interaction among system modules. Testing procedures encompass various stages, including unit testing, integration testing, and user acceptance testing (UAT). Unit testing verifies the functionality of individual components in isolation, while integration testing assesses their interoperability and functionality as a cohesive system. UAT involves stakeholders and clinical users evaluating the platform's usability, performance, and alignment with user requirements, providing valuable feedback for improvement.

Furthermore, quality assurance measures are paramount throughout the integration and testing phase. Data integrity checks are conducted to ensure the accuracy and consistency of stored and processed data, while robust security measures are implemented and tested to safeguard sensitive data and ensure compliance with privacy regulations. Performance testing evaluates metrics such as response time, throughput, and scalability to assess the platform's ability to handle varying loads effectively.

The integration and testing process follows an agile development approach, allowing for flexibility and adaptability to evolving project requirements and stakeholder needs. Feedback from testing phases, including user feedback and performance metrics, is carefully analyzed and incorporated into iterative refinements of the DIOPTRA platform, ensuring continuous improvement and optimization.

Overall, the integration and testing of the DIOPTRA platform are crucial steps in its development lifecycle, ensuring its reliability, functionality, and usability in advancing cancer screening and early detection efforts. Through rigorous testing and iterative refinement, the platform is poised to make a significant impact in the fight against colorectal cancer and beyond.

5 Discussion

The development and implementation of the DIOPTRA platform represent a significant asset in the research field of cancer screening and early detection, particularly in the context of colorectal cancer (CRC). Through the integration of advanced data management and analytics, DIOPTRA offers a comprehensive solution aimed at addressing key challenges in CRC prevention and diagnosis.

One of the primary strengths of the DIOPTRA platform lies in its ability to leverage diverse data sources and integrate them into a unified, standardized format. By collecting and curating data on tissue/liquid biopsy samples, demographic, behavioral, environmental, and medical history information, DIOPTRA generates a homogenized dataset that serves as a valuable resource for risk factor investigation and biomarker analysis. This holistic approach to data integration enables researchers and clinicians to gain deeper insights into the complex interplay of factors related to CRC risk and incidence.

The incorporation of AI into the DIOPTRA platform further enhances its capabilities in CRC screening and risk assessment. AI-powered algorithms analyze blood biomarkers, assess extended risk factor pools, and provide clinical decision support, aiding in the early detection of CRC and the identification of high-risk individuals. By leveraging machine learning techniques, DIOPTRA can achieve higher accuracy and efficiency in detecting abnormal tissue changes, thereby improving patient outcomes and reducing mortality rates via early detection.

Moreover, the user-friendly interfaces provided by the DIOPTRA Front-end streamline the data upload and management process, ensuring seamless interaction between clinical users and the platform for accessing and analyzing clinical data, supported by the anonymization tool that ensures privacy preservation in the uploaded datasets. These features enhance the accessibility and usability of the platform, establishing an interoperable research link between clinicians and technical experts and empowering the former to make informed decisions and optimize patient care through standardized data collection and monitoring.

However, despite its numerous strengths, the DIOPTRA platform also faces several challenges and limitations. One such challenge is the need for continuous validation and refinement of AI algorithms to ensure their accuracy and reliability in real-world clinical settings. Additionally, ensuring compliance with privacy regulations and data security standards remains a priority, given the sensitive nature of patient health information. Furthermore, while the DIOPTRA platform shows significant benefits in the coordinated implementation of multi-center clinical studies, its implementation focuses on the participating clinical sites and the project-specific data ontology for CRC research. A potential broader impact through generalization requires further adjustments, scalability steps and evaluation within long-term prospective studies to assess the platform's effectiveness for structured and cost-effective data collection, curation and integration towards subsequent analysis by researchers.

Ultimately, the DIOPTRA platform represents a comprehensive solution for the coordination of a research-oriented clinical study integrating diverse data sources, AI, and user-friendly interfaces. By leveraging these technologies, DIOPTRA has the potential to improve patient outcomes and reduce CRC mortality rates through early detection, advancing the field of cancer prevention.

6 Conclusion

In conclusion, this paper has provided an overview of the DIOPTRA platform's architectural design and its potential implications for addressing colorectal cancer (CRC)

research challenges within interdisciplinary clinical studies, by leveraging innovative technologies such as artificial intelligence, data integration, and user-friendly interfaces. Through the integration of the DIOPTRA Back-end, which facilitates data ingestion, curation, and storage, with the Front-end's user-friendly interfaces and the use of a dedicated anonymization tool, the platform offers the required link between clinical practice and innovative research in CRC screening programs. Moving forward, it is imperative to continue the integration and testing efforts of the DIOPTRA platform to ensure its effectiveness and scalability in real-world clinical settings.

This paper underscores the importance of ongoing efforts to advance cancer screening technologies and highlights the role of integrated platforms like DIOPTRA in shaping the future of healthcare delivery. By fostering interdisciplinary partnerships and embracing continuous innovation, DIOPTRA has the potential to revolutionize CRC screening practices, reduce mortality rates, and ultimately improve public health outcomes.

Conflict of Interest. The authors declare that they have no conflict of interest.

Acknowledgement. Funded by the European Union (DIOPTRA, 101096649, https://www.dioptra-project.eu/). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

This work has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI).

Funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10056682].

We extend our gratitude to the University of Ioannina (UOI) and the Centre Hospitalier Universitaire de Liège (CHUL) who have undertaken the roles of technical and clinical coordinator, for their contribution and support throughout this project, including the implementation of this paper's architectural design.

References

1. Argilés G, Tabernero J, Labianca R, Hochhauser D, Salazar R, Iveson T, Laurent-Puig P, Quirke P, Yoshino T, Taieb J, Martinelli E, Arnold D (2020) Localised colon cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Annals of oncology : official journal of the European Society for Medical Oncology 31:1291–1305. https://doi.org/10.1016/J.ANNONC.2020.06.022
2. Xi Y, Xu P (2021) Global colorectal cancer burden in 2020 and projections to 2040. Translational Oncology 14:101174. https://doi.org/10.1016/J.TRANON.2021.101174
3. Riihimäki M, Thomsen H, Sundquist K, Sundquist J, Hemminki K (2018) Clinical landscape of cancer metastases. Cancer medicine 7:5534–5542. https://doi.org/10.1002/CAM4.1697
4. Gini A, Jansen EEL, Zielonke N, Meester RGS, Senore C, Anttila A, Segnan N, Mlakar DN, de Koning HJ, Lansdorp-Vogelaar I, Veerus P, Heinävaara S, Sarkeala T, Csanádi M, Pitter

J, Széles G, Vokó Z, Minozzi S, van Ballegooijen M, Driesprong - de Kok I, Heijnsdijk E, Jansen E, van Ravesteyn N, Ivanus U, Jarm K, Primic-Žakelj M, McKee M, Priaulx J (2020) Impact of colorectal cancer screening on cancer-specific mortality in Europe: A systematic review. European journal of cancer (Oxford, England : 1990) 127:224–235. https://doi.org/10.1016/J.EJCA.2019.12.014

5. Marcellinaro R, Spoletini D, Grieco M, Avella P, Cappuccio M, Troiano R, Lisi G, Garbarino GM, Carlini M (2024) Colorectal Cancer: Current Updates and Future Perspectives. Journal of Clinical Medicine 13:40. https://doi.org/10.3390/jcm13010040
6. Yu J, Feng Q, Kim JH, Zhu Y (2022) Combined Effect of Healthy Lifestyle Factors and Risks of Colorectal Adenoma, Colorectal Cancer, and Colorectal Cancer Mortality: Systematic Review and Meta-Analysis. Front Oncol 12:827019. https://doi.org/10.3389/fonc.2022.827019
7. Mosquera I, Mendizabal N, Martín U, Bacigalupe A, Aldasoro E, Portillo I, from the Desberdinak Group (2020) Inequalities in participation in colorectal cancer screening programmes: a systematic review. European Journal of Public Health 30:558–567. https://doi.org/10.1093/eurpub/ckz236
8. Li S, Miller-Wilson L-A, Guo H, Fisher DA (2022) Adherence to colorectal cancer screening and healthcare resource utilization: a longitudinal analysis in Medicare beneficiaries aged 66-75 years. Curr Med Res Opin 38:2201–2208. https://doi.org/10.1080/03007995.2022.2133493
9. Bénard F, Barkun AN, Martel M, von Renteln D (2018) Systematic review of colorectal cancer screening guidelines for average-risk adults: Summarizing the current global recommendations. World J Gastroenterol 24:124–138. https://doi.org/10.3748/wjg.v24.i1.124
10. Wang L, Lo C-H, He X, Hang D, Wang M, Wu K, Chan AT, Ogino S, Giovannucci EL, Song M (2020) Risk Factor Profiles Differ for Cancers of Different Regions of the Colorectum. Gastroenterology 159:241-256.e13. https://doi.org/10.1053/j.gastro.2020.03.054
11. Mazouji O, Ouhajjou A, Incitti R, Mansour H (2021) Updates on Clinical Use of Liquid Biopsy in Colorectal Cancer Screening, Diagnosis, Follow-Up, and Treatment Guidance. Frontiers in Cell and Developmental Biology 9:
12. Akimoto N, Ugai T, Zhong R, Hamada T, Fujiyoshi K, Giannakis M, Wu K, Cao Y, Ng K, Ogino S (2021) Rising incidence of early-onset colorectal cancer: a call for action. Nat Rev Clin Oncol 18:230–243. https://doi.org/10.1038/s41571-020-00445-1
13. Spychalski P, Kobiela J, Wieszczy P, Bugajski M, Reguła J, Kaminski MF (2021) Adenoma to Colorectal Cancer Estimated Transition Rates Stratified by BMI Categories—A Cross-Sectional Analysis of Asymptomatic Individuals from Screening Colonoscopy Program. Cancers (Basel) 14:62. https://doi.org/10.3390/cancers14010062
14. Slijepčević D, Henzl M, Klausner LD, Dam T, Kieseberg P, Zeppelzauer M (2021) k-Anonymity in practice: How generalisation and suppression affect machine learning classifiers. Computers & Security 111:102488. https://doi.org/10.1016/j.cose.2021.102488
15. Karacapilidis N, Tsakalidis D, Domalis G (2023) An AI-Enhanced Solution for Large-Scale Deliberation Mapping and Explainable Reasoning. In: Papadaki M, Rupino da Cunha P, Themistocleous M, Christodoulou K (eds) Information Systems. Springer Nature Switzerland, Cham, pp 305–316