ProvideQ: A Web-Based Knowledge Platform for Assessing Preanalytical Influences on Biomolecules in Biospecimens

Sven Bichtemann, Oliver J. Bott, Johanna Apfel-Starke, Vicky Scholz, Thomas Illig, Sara Haag

Suggested citation:

Bichtemann, Sven, Oliver J. Bott, Johanna Apfel-Starke, Vicky Scholz, Thomas Illig, and Sara Haag. 2025. "ProvideQ: A Web-Based Knowledge Platform for Assessing Preanalytical Influences on Biomolecules in Biospecimens." In *German Medical Data Sciences 2025: GMDS Illuminates Health (Studies in Health Technology and Informatics; 331)*, 386–94. IOS Press. https://doi.org/10.25968/opus-3807.

Abstract

Introduction: Preanalytical factors significantly impact the stability of biomolecules in biospecimens, affecting the reliability of biomedical research and diagnostics. This paper presents the development of ProvideQ (Database for pre-analytical variability and biospecimen quality), a web-based platform designed to centralize access to research findings on these influences.

Methods: Building on an initial prototype, we implemented a validated criteria catalog for data quality, an efficient search system handling incomplete inputs, and SPREC 4.0 integration for standardized coding of preanalytical variables. User feedback from usability tests enhanced the platform's interface.

Results: Results include an improved data model, a Python-based literature import module, and an intuitive frontend using Next.js and React.

Conclusion: ProvideQ supports analyte- and sample-centric searches, demonstrating its potential as a valuable tool in biobanking and research. Future enhancements include expanding the database and integrating AI-driven analytics of scientific publications on preanalytical factors to facilitate the import of research results into the platform.

German Medical Data Sciences 2025: GMDS Illuminates Health
R. Röhrig et al. (Eds.)

doi:10.3233/SHTI251419

ProvideQ: A Web-Based Knowledge Platform for Assessing Preanalytical Influences on Biomolecules in Biospecimens

Sven BICHTEMANN[a,1], Oliver Johannes BOTT[b], Johanna APFEL-STARKE[b,c], Vicky SCHOLZ[a], Thomas ILLIG[a], and Sara HAAG[a,d]
[a] *Hannover Unified Biobank (HUB), Medizinische Hochschule Hannover, Carl Neuberg Str.1, 30625 Hannover*
[b] *Institute for Applied Data Science Hannover DATA|H, University of Applied Sciences and Arts, Hannover, Germany*
[c] *MHH Information Technology, Medizinische Hochschule Hannover, Hannover*
[d] *Translationsallianz in Niedersachsen (TRAIN), Feodor-Lynen-Straße 7, 30625 Hannover*
ORCiD ID: SB 0009-0006-4746-5608 , OJB 0000-0002-8823-5247
JAS 0000-0001-7068-1508 , VS 0009-0008-1852-444X
TI 0000-0003-4284-5389 SH 0000-0003-4284-5389

Abstract. Introduction: Preanalytical factors significantly impact the stability of biomolecules in biospecimens, affecting the reliability of biomedical research and diagnostics. This paper presents the development of ProvideQ (Database for pre-analytical variability and biospecimen quality), a web-based platform designed to centralize access to research findings on these influences. **Methods:** Building on an initial prototype, we implemented a validated criteria catalog for data quality, an efficient search system handling incomplete inputs, and SPREC 4.0 integration for standardized coding of preanalytical variables. User feedback from usability tests enhanced the platform's interface. **Results:** Results include an improved data model, a Python-based literature import module, and an intuitive frontend using Next.js and React. **Conclusion:** ProvideQ supports analyte- and sample-centric searches, demonstrating its potential as a valuable tool in biobanking and research. Future enhancements include expanding the database and integrating AI-driven analytics of scientific publications on pre-analytical factors to facilitate the import of research results into the platform.

Keywords. Pre-Analytical Phase, Specimen Handling, Biological Specimen Banks, Quality Control

[1] Corresponding author: Sara Haag, Hannover Unified Biobank (HUB), Medizinische Hochschule Hannover, Carl Neuberg Str.1, 30625 Hannovor, Translationsallianz in Niedersachsen (TRAIN), Feodor-Lynen-Straße 7, 30625 Hannover, haag.sara@mh-hannover.de..

1. Introduction

The integrity of biomolecules - such as metabolites, proteins, or RNA - in biospecimens is a necessary prerequisite for the validity of biomedical research and clinical diagnostics based on these samples. However, preanalytical variables, including time-to-centrifugation, storage temperature, and sample handling, can compromise this integrity, leading to variability that undermines result reproducibility [1]. Standardizing these conditions is a persistent challenge, addressed in part by the Standard PREanalytical Code (SPREC), which systematically documents key preanalytical parameters [2]. Despite such efforts, a centralized, accessible resource linking these factors to specific biomolecule stability data remains absent, leaving researchers and biobank personnel reliant on fragmented literature searches.

The ProvideQ platform emerged from this gap, initiated as a prototype in a prior project at Hannover Medical School's Hannover Unified Biobank (HUB) and the University of Applied Sciences and Arts Hannover (HsH).

This project addresses a key question in the field of biobanking quality research: whether it is feasible to develop a database system capable of capturing the often highly heterogeneous literature concerning the impact of pre-analytical factors on biospecimens. The aim is to enable users to efficiently and clearly identify which specific pre-analytical conditions affect particular types of biospecimens and to what extent, as well as how consistent or contradictory literature is regarding certain analytes or specimen types. Ultimately, the project seeks to assess to what extent such structured information can support improved decision-making, for example, in study planning or in determining whether existing samples, exposed to specific conditions, may still be used to generate reproducible and high-quality results, thereby enhancing the overall quality of biospecimen-based biomedical research.

The relevance and potential of the ProvideQ database approach were underscored by a poster prize awarded to the ProvideQ concept at the 12th National Biobank Symposium [3, [4]. This work advances that foundation into a functional web-based platform.

The technical objective of this development, as presented in this paper, was to evolve ProvideQ into a fully operational and user-friendly platform that allows for the exploration of preanalytical influences on biomolecules based on relevant scientific literature. This endeavor aimed to answer several key scientific and technical questions: (1) How can a criteria catalog be designed and validated to ensure the quality and consistency of data extracted from literature, particularly when SPREC parameters are incompletely documented? (2) What algorithms and system architecture are suitable for an efficient search system that can process complex queries with partial or missing parameters and accurately assign analytes? (3) How can the SPREC 4.0 standard be reliably integrated into the platform to ensure accurate encoding and decoding of preanalytical variables, thereby promoting standardization? (4) Which user-centered evaluation methods are most effective for iteratively refining the platform's usability and functionality to meet the needs of researchers and biobank staff? This paper details the development process, technical implementation, and initial evaluation addressing these questions, positioning ProvideQ as a pivotal tool for biobanking and research.

2. Methods

Development began with an analysis of the prototype's state, comprising a PostgreSQL database, a Flask-based backend, and a Next.js frontend [3]. Key challenges were identified: unstructured literature integration, limited search capabilities, outdated SPREC integration (version 3.0), and untested usability. The approach, grounded in agile methodology, unfolded iteratively through the phases of requirements analysis, system design, implementation, and testing.

2.1. Requirement Analysis and Data Model

Two primary user groups were defined using the personas approach: researchers seeking analyte-specific stability data and biobank staff assessing sample quality. A key challenge identified during requirements analysis was the difficulty in establishing a unique and consistent nomenclature for metabolites, given the frequent use of synonyms and variations across scientific publications. Requirements therefore included a robust import mechanism for data concerning preanalytical influences documented in scientific publications, flexible search accommodating partial data, and SPREC 4.0 compatibility. To address the metabolite naming challenge and ensure standardized analyte identification, the requirements mandated the integration of established reference databases; consequently, the system concept included incorporating HMDB [5] and RefMet [6] as authoritative sources and templates for uniform naming. Furthermore, the analysis highlighted the significant complexity and heterogeneity of published research findings on preanalytical factors, often featuring incomplete documentation of experimental conditions. A criteria catalog was developed to ensure data consistency and enable robust search functionality. This catalog, implemented as part of the CSV import validation logic, mandates the inclusion of key data fields for each literature entry: analyte name (or RefMet identifier), Digital Object Identifier (DOI) of the source publication, method name, method type, sample preparation details, and the primary sample type (e.g., serum, plasma). Preanalytical factors (e.g., centrifugation time, storage temperature) were defined as optional fields within the preanalytical_factor database table to accommodate the common incompleteness of such details in published studies, allowing for flexible data integration while ensuring core traceability and analyte-method-sample context.

2.2. Implementation

The platform utilizes a three-tier architecture, a common software design pattern chosen for its modularity, scalability, and maintainability, clearly separating presentation (frontend), application logic (backend), and data storage (database).

Backend (Flask [7]): A RESTful API was built with endpoints for literature import (/api/v1/literature) and search. A Python script automates CSV-based data imports. This script reads CSV files, validates each row against the mandatory fields defined in the criteria catalog, and performs data type conversions. It then identifies or creates entries for the source literature (based on DOI), analytical method, preanalytical conditions, and documented effects, linking them to the corresponding analyte (matched against analyte, synonymous, or refmet tables) or RefMet identifier. All database operations within a single CSV row import are handled within a transaction to ensure data integrity.

Frontend (Next.js/React [8]): Pages for analyte-centric (analyte-search/page.js) and sample-centric (sample-search/page.js) searches were designed with Tailwind CSS, a utility-first CSS framework that provides a set of pre-defined, low-level CSS classes [9]. These interfaces provide users with structured input capabilities: (1) an autocomplete-enabled field for analyte selection (validating against analyte, synonymous, and refmet tables); (2) an optional SPREC 4.0 code input section that populates preanalytical factor fields upon SPREC code selection; and (3) a detailed, collapsible section for manual entry or refinement of individual preanalytical parameters (e.g., temperature ranges, time durations). Search results are initially displayed in a CSV-like format directly on the search page, with an option to view them in an interactive, sortable tabular format on a separate results page.

Database (PostgreSQL [10]): Docker-containerized, it supports ~40,000 HMDB [4] and ~180,000 RefMet [6] metabolites, updated to SPREC 4.0, and stores literature-derived effects. The preanalytical_factor table within the database is designed to capture a granular level of detail, including specific numeric values and ranges for various preanalytical steps (e.g., cent_time_from_in_min, pre_cent_temp_to). This structure allows for storing information more detailed than what SPREC level codes might generally represent. While SPREC codes can be used for broader categorization and input, the system stores and searches on these more specific, optionally provided parameters, ensuring that detailed experimental conditions from literature can be preserved and queried.

Search logic was specifically engineered to accommodate the heterogeneity of data in scientific literature and user query patterns. The backend API endpoint processes queries by dynamically constructing SQL WHERE clauses. A key feature is the handling of range-based parameters (e.g., pre-centrifugation temperature 19–22°C). The system retrieves records where the preanalytical parameter ranges stored in the database fall entirely within or are identical to the user-specified ranges. This ‘database-range-within-user-range’ logic ensures that studies reporting precise measurement points are found when users query broader SPREC-defined categories or custom ranges. Furthermore, if users provide only one bound of a range (e.g., ‘temperature from 19°C’ without a ‘to’ value), the search adapts to filter only on the provided bound. This flexibility allows for meaningful results even with incomplete or partial preanalytical factor inputs. For sample-centric searches, typically initiated by biobank staff to assess overall sample suitability based on SPREC codes or specific preanalytical conditions without pre-selecting an analyte, the system queries all analyte records matching the provided preanalytical criteria, returning a list of affected biomolecules and the corresponding literature evidence

2.3. Evaluation

The platform’s evaluation combined technical validation and user feedback. A qualitative usability test was conducted with one biobank staff member, representing a key target user group. This session employed a task-based, think-aloud protocol where the participant performed predefined search tasks for both analyte-centric and sample-centric use cases. Observations were recorded, and a custom-developed, structured questionnaire was used to gather specific feedback on navigation, search functionality, result presentation, and overall ease of use. A standardized questionnaire (e.g., System Usability Scale) was not employed at this stage due to the prototype’s early development phase and the need for specific feedback on newly implemented core functionalities

tailored to ProvideQ's unique use cases. The custom questionnaire was designed to elicit detailed qualitative insights into these specific areas. User feedback led to targeted interface enhancements, including collapsible text sections for improved information hierarchy, a fixed table header for better scannability of large result sets, and options for CSV export of results and analyte lists.

Backend stability was confirmed through unit tests and integration tests covering API endpoint functionality and database interactions. Functional validation employed practical use cases simulating researcher (analyte-centric) and biobank staff (sample-centric/SPREC-based) queries. These tests verified accurate handling of synonym ambiguity (e.g., "2-Hydroxybutyric acid"), precise filtering based on preanalytical parameters (including ranges), and correct retrieval of data from the integrated literature. While the usability evaluation was limited to a single participant, the detailed qualitative feedback proved valuable for iterative refinement of the user interface and prioritization of features. Results include an improved data model, a Python-based literature import module, and an intuitive frontend using Next.js and React.

3. Results

Technical implementation succeeded in integrating a scalable infrastructure. The import module successfully processed 1,344 literature entries derived from a single, extensively evaluated publication (Kamlage et al., 2013) [6], resulting in the integration of 435 unique records and 111 additional records based on synonym matches. 909 entries were skipped due to unidentifiable analytes in the current reference datasets (HMDB, RefMet). Search functionality demonstrably supports both analyte- and sample-centric queries. Analyte-centric searches were validated using test cases, such as for "2-Hydroxybutyric acid" with precentrifugation conditions of 19–22°C, which correctly returned the single matching record. Sample-centric queries, demonstrated with a SPREC code (e.g., BLD-SED-A1-C- -- -E-A), successfully retrieved multiple relevant analytes, confirming the system's capability to identify all biomolecules affected by specific preanalytical conditions from the imported dataset.

In the frontend, autocomplete functionalities for analyte names (drawing from analyte, synonymous, and RefMet tables) effectively reduce input errors, while SPREC 4.0 dropdowns simplify preanalytical parameter entry. Initially, results were presented in a CSV-like format; however, based on user feedback, the interface was enhanced to include a dedicated tabular view with expandable rows for detailed information, options for CSV export of all or selected data, and an analyte list export for sample-centric search results. Features such as collapsible information sections and a fixed table header further improved usability.

Figure 1 illustrates the user workflow in ProvideQ: specifying an analyte (2-Hydroxybutyric acid) and preanalytical conditions (centrifugation parameters) yields a tabular result. The output columns detail bibliographic information (DOI, author, year), sample characteristics, and recorded preanalytical effects, centrally captured in the 'Concentration change' variable (e.g., 'no significant change', 'significant increase'), alongside statistical significance indicators (p-value, FDR). The highlighted row and framed cell in the example emphasize a finding of instability ('significant increase'), allowing users to quickly identify critical results derived from the literature.

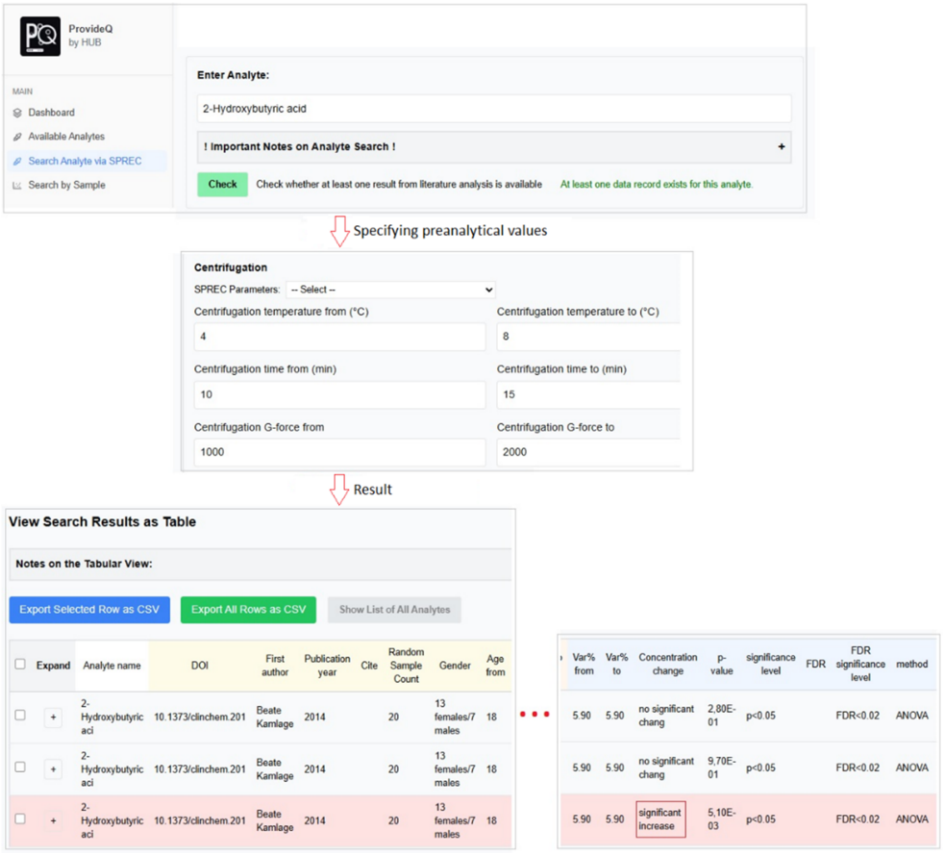

Figure 1: Search process to find entries for the analyte 2-Hydroxybutyric acid under specific centrifugation conditions.

4. Discussion

Researchers and biobank staff working with biosamples require reliable information on any sample preparation and storage aspects that may influence the analyses of the samples. To the best of our knowledge, a web-based information system for pre-analytical influencing factors does not yet exist. ProvideQ fills this gap and advances biobanking by consolidating preanalytical stability data into an accessible, searchable platform. The developed prototype successfully addresses the initial challenges by implementing a flexible search logic, a validated data import mechanism, and an enhanced user interface based on initial feedback.

A central challenge in this domain is the heterogeneity and frequent incompleteness of data in scientific publications. The implemented database schema and search logic address this by design. The preanalytical_factor table allows for the storage of specific, granular data points (e.g., precise centrifugation times or temperatures) that go beyond the broader categories of SPREC codes. This ensures that more detailed information from the literature beyond SPREC does not have to be simplified or is not lost during

entry. The search function's 'database-range-within-user-range' logic further supports this flexibility, enabling users to find relevant studies even if their query parameters are broader than the specific data points in the database. When encountering conflicting results from different studies - for example, where one paper reports an analyte as stable under certain conditions while another reports degradation - the current implementation transparently displays all retrieved findings. This allows users to see the complete evidence base, though future iterations could implement visualization or ranking mechanisms to better contextualize such discrepancies.

However, several limitations must be acknowledged. The usability evaluation, while providing valuable qualitative insights that led to concrete UI improvements, was restricted to a single expert user. This limits the generalizability of the findings, and a broader evaluation with a diverse group of researchers and biobank staff from multiple institutions is a critical next step to ensure the tool's broader utility and intuitiveness. Furthermore, the platform's database is currently limited, primarily containing data from one comprehensively annotated publication alongside reference data from HMDB and RefMet. For ProvideQ to become an indispensable tool, a significant expansion of this knowledge base is required.

The scalability of data acquisition is therefore the most significant future challenge. While the current CSV import is more scalable than manual entry, it remains a bottleneck. To overcome this, we envision a semi-automated pipeline leveraging AI and Large Language Models (LLMs). This future methodology would involve developing models to automatically extract relevant information - such as analytes, preanalytical parameters, and outcomes - from unstructured text and tables in scientific publications. Such a system could be integrated with literature databases like PubMed or Web of Science to automatically screen new publications and suggest structured data for import, drastically reducing manual effort and enabling exponential growth of the database. The modular three-tier architecture of ProvideQ was intentionally chosen to support such future extensions and ensure the platform can scale to accommodate a growing dataset and user base. Securing long-term funding through research grants or academic-industrial partnerships will be essential to support these scaling efforts, including development, data curation, and maintenance.

5. Conclusion

ProvideQ has been successfully developed as a functional prototype, demonstrating its feasibility as a specialized knowledge platform for preanalytical research. The work presented here yields several key insights essential for the development and scaling of such systems:

- A flexible data model, which treats specific preanalytical parameters as optional while enforcing core metadata (e.g., DOI, sample type), is crucial for accommodating the heterogeneous and often incomplete data found in scientific literature.
- The implemented 'database-range-within-user-range' search logic proves effective for handling complex queries, allowing users to find relevant results from specific data points even when querying broader parameter ranges.

- Automated, criteria-based data import mechanisms are vital for the scalability of the knowledge base, proving far more efficient and consistent for processing large volumes of literature-derived data than manual entry.
- Even limited, task-based user feedback is invaluable for driving targeted functional enhancements, such as improved result visualization and data export features, that directly align the platform with the practical workflows of its target users.

These findings underscore that ProvideQ is not only a functional tool but also a model for future developments in this domain. Its modular design supports scalability, with planned enhancements including AI-driven literature analysis and public deployment. A longer-term vision of the project is to leverage this structured data to potentially identify novel biomarker patterns for biospecimen quality using artificial intelligence approaches. These steps will be critical in positioning ProvideQ as a standard resource that enhances research reproducibility and biobank quality assurance.

Declarations

Conflict of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contributions: SB: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Visualization, Writing – original draft. OJB: Conceptualization, Methodology, Supervision, Writing – review & editing. JAS: Conceptualization, Methodology, Supervision, Writing – review & editing. VS: Validation, Resources, Writing – review & editing. TI: Conceptualization, Supervision, Resources, Funding acquisition, Writing – review & editing. SH: Conceptualization, Supervision, Project administration, Resources, Funding acquisition, Writing – review & editing.

Acknowledgement: We would like to thank the Hannover Unified Biobank (HUB) team for their valuable support with questions about biobanking and for the helpful user feedback. Special thanks go to Burak Bozkurt for his support in preparing the Docker environment. We would also like to thank the ProvideQ team at Hochschule Hannover for their excellent cooperation during the initial preparatory work.

References

[1] Lippi G, von Meyer A, Cadamuro J, Simundic AM. Blood sample quality [Internet]. Vol. 6, Diagnosis. Walter de Gruyter GmbH; 2018. p. 25–31. [cited 19 Apr 2025]. Available from: http://dx.doi.org/10.1515/dx-2018-0018.

[2] Betsou F, Chuaqui R, De-Wilde A, Grizzle W, Kiehntopf M, Linsen L, et al. Standard PREanalytical Code Version 4.0 [Internet]. Biopreservation and Biobanking. Mary Ann Liebert Inc; 2024. [cited 2025 Apr 19]. Available from: http://dx.doi.org/10.1089/bio.2024.0010.

[3] Haucke M, Huei S-Y, Wiesner J, Badorek S, Bichtemann S, Berlage W, Bozkurt Burak, Paszkier P, Polischuk A, Dembowski L, Rabbe L, Rezaei E, Bott OJ, Apfel-Starke J, Illig T, Haag S: ProvideQ: Datenbank für prä-analytische Variabilität und Bioprobenqualität. In: Altmann H, Anton G, Baber R et al. (Hrsg.): Vernetztes Biobanking: Gemeinsam stark in die Zukunft. Proceedings of the 12. National Biobank Symposium (NBS 2024), 23.–24. September 2024 in Berlin, ibidem-Verlag, 2024, S. 189–191.

[4] Medizinische Hochschule Hannover. HUB erhält Posterpreis auf dem NBS 2024 [Internet]. 2024. [cited 2025 Jun 27]. Available from: https://www.mhh.de/institute-zentren-forschungseinrichtungen/hannover-unified-biobank-hub/aktuelles/standard-titel-13.
[5] Wishart DS, Guo A, Oler E, Wang F, Anjum A, Peters H, et al. HMDB 5.0: the Human Metabolome Database for 2022 [Internet]. Vol. 50, Nucleic Acids Research. Oxford University Press (OUP); 2021. p. D622–31. [cited 2025 Apr 20]. Available from: http://dx.doi.org/10.1093/nar/gkab1062.
[6] Fahy E, Subramaniam S. RefMet: a reference nomenclature for metabolomics [Internet]. Vol. 17, Nature Methods. Springer Science and Business Media LLC; 2020. p. 1173–4. [cited 2025 Apr 20]. Available from: http://dx.doi.org/10.1038/s41592-020-01009-y.
[7] Welcome to flask — flask documentation (3.1.X) [Internet]. Palletsprojects.com. [cited 2025 Jun 28]. Available from: https://flask.palletsprojects.com/en/stable/.
[8] Next.Js by Vercel - the React framework [Internet]. Nextjs.org. [cited 2025 Jun 28]. Available from: https://nextjs.org/.
[9] Rapidly build modern websites without ever leaving your HTML [Internet]. Tailwindcss.com. [cited 2025 Jun 28]. Available from: https://tailwindcss.com/.
[10] PostgreSQL [Internet]. PostgreSQL. 2025 [cited 2025 Jun 28]. Available from: https://www.postgresql.org/.
[11] Kamlage B, Maldonado SG, Bethan B, Peter E, Schmitz O, Liebenberg V, et al. Quality Markers Addressing Preanalytical Variations of Blood and Plasma Processing Identified by Broad and Targeted Metabolite Profiling [Internet]. Vol. 60, Clinical Chemistry. Oxford University Press (OUP); 2014. p. 399–412. [cited 21 Apr 2025]. Available from: http://dx.doi.org/10.1373/clinchem.2013.211979.