BioASQ at CLEF2025: The thirteenth edition of the large-scale biomedical semantic indexing and question answering challenge

Anastasios Nentidis[1], Georgios Katsimpras[1], Anastasia Krithara[1], Martin Krallinger[2], Miguel Rodriguez Ortega[2], Natalia Loukachevitch[3], Andrey Sakhovskiy[4,6], Elena Tutubalina[4,5,6], Grigorios Tsoumakas[7], George Giannakoulas[7], Alexandra Bekiaridou[8], Athanasios Samaras[7], Giorgio Maria Di Nunzio[9], Nicola Ferro[9], Stefano Marchesin[9], Laura Menotti[9], Gianmaria Silvello[9], and Georgios Paliouras[1]

[1] National Center for Scientific Research "Demokritos", Athens, Greece
{tasosnent, gkatsibras, akrithara, paliourg}@iit.demokritos.gr
[2] Barcelona Supercomputing Center, Barcelona, Spain
{martin.krallinger, miguel.rodriguez}@bsc.es
[3] Moscow State University, Russia
louk_nat@mail.ru
[4] Sber AI, Russia
[5] Artificial Intelligence Research Institute, Russia
[6] Kazan Federal University, Russia
{andrey.sakhovskiy, tutubalinaev}@gmail.com
[7] Aristotle University of Thessaloniki, Greece
greg@csd.auth.gr, {g.giannakoulas, th.samaras.as}@gmail.com
[8] Northwell Health, USA
ampekiaridou@gmail.com
[9] University of Padua, Italy
{dinunzio, ferro, stefano.marchesin, laura.menotti, silvello}@dei.unipd.it

Abstract. During the last twelve years, the large-scale biomedical semantic indexing and question-answering challenge (BioASQ) has been pushing towards the continuous advancement of methods and tools to accelerate access to the ever-increasing scientific resources of the biomedical domain. In this direction, each year, BioASQ organizes shared tasks representing the real information needs of biomedical experts and provides respective benchmark datasets. This way, it provides a unique common testbed where research teams around the world can test and compare new approaches for accessing biomedical knowledge.

The thirteenth version of BioASQ will be held as an evaluation Lab in the context of CLEF2025 providing six tasks: (i) *Task b* on biomedical semantic question answering. (ii) *Task Synergy* on question answering developing biomedical topics. (iii) *Task MultiClinSum* on multilingual clinical summarization. (iv) *Task BioNNE-L* on nested named entity linking in Russian and English. (v) *Task ELCardioCC* on clinical coding in cardiology. (vi) *Task GutBrainIE* on gut-brain interplay information extraction. As BioASQ rewards the methods that outperform the state of

the art in these shared tasks, it keeps pushing the research frontier towards approaches that will meet the need for efficient and precise access to biomedical knowledge.

Keywords: Biomedical · Semantic Indexing · Question Answering

1 Introduction

BioASQ[10] [27] is a series of international challenges and workshops on biomedical semantic indexing and question answering (QA) introduced in 2012. Each year, BioASQ is structured into distinct shared tasks relevant to biomedical information access, machine learning, information retrieval, information extraction, and others, and a BioASQ workshop is also organized [23, 21, 22, 6, 3]. Unique benchmark datasets and open-source infrastructure are developed for running the BioASQ tasks, allowing research teams that work on biomedical information access systems, to compete in the same realistic benchmark datasets and share, evaluate, and compare their ideas and approaches. So far, more than 100 teams from 32 countries have participated in BioASQ. As BioASQ keeps rewarding the most successful approaches in each task and sub-task, it eventually pushes toward systems that outperform previous approaches. Such successful approaches for semantic indexing and QA can eventually lead to the development of tools to support more precise access to biomedical knowledge and to further improve health services.

2 BioASQ evaluation lab 2025

The thirteenth BioASQ challenge (BioASQ13) will be part of the sixteenth CLEF conference[11] and will consist of six tasks that are central to biomedical knowledge access and the question-answering process: (i) *Task b*[12] on the processing of biomedical questions, the generation of answers, and the retrieval of supporting material, (ii) *Task Synergy* on biomedical QA for developing problems under a scenario that promotes collaboration between biomedical experts and question-answering systems, (iii) *Task MultiClinSum* on Multilingual Clinical Summarization. (iv) *Task BioNNE-L* on the automated concept normalization of nested named entities from documents written in Russian and English. (v) *Task ELCardioCC* on clinical coding of Greek cardiology discharge letters. (vi) *Task GutBrainIE* on extracting and linking knowledge from the scientific literature on the gut-brain interplay. As *Task b* and *Task Synergy* have also been organized in the context of previous editions of the BioASQ challenge [14, 13, 18], we refer to their current version, in the context of BioASQ13, as *task 13b*, *task Synergy 13* respectively.

[10] http://www.bioasq.org

[11] http://clef2025.clef-initiative.eu/

[12] Since the first BioASQ, the task on biomedical semantic indexing has been called *Task a* and the task on QA *Task b*, for brevity. *Task a* was completed in 2020 [4].

2.1 Task 13b: Biomedical question answering

BioASQ *task 13b* takes place in three phases. In Phase A, biomedical questions in English are provided and the systems have to retrieve relevant material (PubMed documents and snippets). In Phase A+, the participating systems have to provide 'exact' and 'ideal' answers. Depending on question type, the 'exact' answer can be a *yes* or *no* (yes/no), an entity name, such as a disease or gene (factoid), or a list of entity names (list). The 'ideal' answer is a paragraph-sized summary, regardless of question type. Finally, in Phase B, some relevant material is provided for each question, selected by the BoASQ experts, and the systems have to provide new answers given this additional information.

About 300 new biomedical questions annotated with golden documents, snippets, and answers ('exact' and 'ideal'), will be developed for testing. In addition, a training set of about 5,380 biomedical questions, accompanied by answers, and supporting evidence (documents and snippets), will be available from previous versions of the tasks, as a unique resource for the development of question-answering systems [5]. The evaluation in *task 12b* is done manually by the experts that assess each system response and automatically by employing a variety of established evaluation measures [11] as in the previous version of the task [14].

2.2 Task Synergy 13: Question answering for developing topics

Aiming to promote research in developing biomedical topics, such as COVID-19, we introduced the BioASQ *task Synergy* in 2020 [6, 7]. Contrary to the original *task b*, *task Synergy* is designed as a continuous dialog, that allows the experts to pose open questions for developing topics, for which they do not know beforehand whether a definite answer can be provided. The systems provide relevant material (documents and snippets) and answers and the experts assess them, providing feedback that allows the systems to improve their responses to these questions. This process repeats iteratively and new material is also considered in each round, based on updates of the original document resource [17]. Since 2023, this evolving document resource is PubMed [13].

A training dataset of about 370 questions on developing topics with incremental annotations with relevant material and answers is already available from previous versions of *task Synergy* [16, 15, 20, 19]. During the *task Synergy 13* this set will be extended with more than fifty new open questions on developing health topics. Meanwhile, any existing questions that remain relevant may be enriched with more updated answers and more recent evidence. Tn *task Synergy 13* we use the same evaluation measures with *task 13b*, considering only new material for the information retrieval part, an approach known as *residual collection evaluation*[26]. However, the focus of this task is to aid the experts in contributing to the incremental understanding of new developing health topics and the discovery of new solutions.

2.3 Task MultiClinSum: Multilingual Clinical Summarization

There is a rapid accumulation of different types of clinical content, including medical records and clinical case reports. These are generally written not only in English but actually in a variety of languages. Some of the clinical reports can be very long and thus it is challenging for domain experts to read and keep track of key clinical insights. Large Language Models (LLMs) have shown promising results for summarization approaches that can be useful to condense lengthy clinical cases into a shorter version of text, reducing the size of the initial text while preserving key clinical informational elements. Thus there is a pressing need to evaluate and benchmark how well clinical summarization works for case reports written in different languages.

We introduce the *MultiClinSum* task covering the automatic summarization of lengthy clinical case reports written in different languages, namely English, Spanish, French, and Portuguese. The *MultiClinSum* task will rely on a corpus of manually selected full clinical case reports and their corresponding clinical case report summaries derived from case report publications written in the previously mentioned languages. For evaluation proposes, automatically generated summaries will be compared against manually generated summaries generated by the original authors, exploring Rouge-2 scores and BERTScore [28] for evaluation assessment. As clinical case reports do share commonalities with medical discharge summaries, insights provided by the *MultiClinSum* results can be of practical relevance also for clinical records summarization scenarios.

2.4 Task BioNNE-L: Nested Named Entity Linking in Russian and English

The *BioNNE-L* shared task focuses on NLP challenges in entity linking, also known as medical concept normalization (MCN), for English and Russian languages. The goal is to map biomedical entity mentions to a comprehensive set of medical concept names and their concept unique identifiers (CUIs) from the UMLS. The train/dev datasets include annotated mentions of disorders, anatomical structures, and chemicals. We design our data to account for a complex structure of nested entity mentions and the partial nature of medical terminology. Participants are allowed to train any model architecture on any publicly available data to achieve the best performance. Similar to the *BioNNE 2024 task* [1], the evaluation framework is divided as follows: 1. Track Language-oriented: Participants in this track must develop a model for linking entity mentions in a target language (English or Russian). 2. Track Bilingual: Participants in this track must train a single model using training data for both Russian and English languages. The datasets for task *BioNNE-L* will be based on the NEREL-BIO dataset which includes annotated mentions of disorders, anatomical structures, chemicals, diagnostic procedures, and biological functions[10].

2.5 Task ELCardioCC: Clinical Coding in Cardiology

Cardiovascular diseases affect a significant portion of the global population, accounting for 32% of global deaths according to WHO. Automated clinical coding plays a crucial role in transforming unstructured real-world medical data gathered from patients into structured information, in order to facilitate clinical research and analysis. However, existing research predominantly focuses on English clinical text, leaving other languages, such as Greek, underrepresented. To this end, we propose a new *ELCardioCC* task, which concerns i) the assignment of cardiology-related ICD-10 codes to discharge letters from Greek hospitals, ii) the extraction of the specific mentions of ICD-10 codes from the discharge letters. We will use evaluation metrics, such as micro and macro F-measure for subtask (i) and token F-measure for subtask (ii). The *ELCardioCC* dataset includes 500 cardiology discharge letters in Greek, annotated by at least two experts at the document level with ICD-10 codes, as well as at the corresponding mention level with the ICD-10 code mentions. The dataset annotations include more than 200 ICD-10 codes and 5000 text spans. The dataset includes 5 types of entities: chief complaint, diagnosis, prior medical history, drugs, and cardiac echo.

2.6 Task GutBrainIE: Gut-Brain interplay Information Extraction

Recent evidence suggests a connection between neurological and gut disorders that may play a critical role in mental health-related disorders or diseases like Multiple Sclerosis, Parkinson's, and Alzheimer's. The *GutBrainIE* task aims to foster the development of Information Extraction (IE) systems that support experts by automatically extracting and linking knowledge from scientific literature, facilitating the understanding of gut-brain interplay and its role in neurological diseases. The task is divided into two subtasks. In the first subtask, participants are provided with PubMed abstracts discussing the gut-brain interplay and asked to extract named entities about the gut-brain interplay from PubMed abstracts to link them to the corresponding concepts in a reference ontology. In the second subtask, the participants are asked to identify binary relations – i.e., presence/absence – between any pair of entities they extract within an abstract. The submitted runs are evaluated based on Precision, Recall, and F1 measures for each subtask using gold annotations created by domain experts.

The dataset for the GutBrainIE task includes circa 1,000 PubMed abstracts annotated by experts with entity mentions, corresponding concepts in the reference ontology, and binary relations. The dataset includes various types of entities, such as genes, bacteria, intermediates, and diseases, and binary relationships between them. At the task's start, participants will be provided with the train and validation (dev) datasets, while the test set will become available a couple of weeks before the submission deadline. Together with the dataset, we provide the reference ontology to link the extracted named entities to the corresponding concepts.

2.7 BioASQ datasets and tools

During the twelve years of BioASQ, hundreds of systems from research teams around the world have been evaluated on tasks related to biomedical information access. In this direction, BioASQ has developed a lively ecosystem of tools[13], such as the BioASQ Annotation Tool [24] for question-answering dataset development and a range of evaluation measures [11] and datasets[14], which are publicly available. Beyond the unique datasets provided for the six tasks running this year, BioASQ also provides: i) a benchmark dataset of more than 16.2 million articles on biomedical semantic indexing (task*task a*) [4], ii) the *task MESINESP* datasets [2, 25] of more than 300K articles, on medical semantic indexing in Spanish, iii) the *task DisTEMIST* [12], *task MedProcNER* [8], and *task MultiCardioNER* [9] on medical information extraction from clinical case documents, iv) the *task BioNNE task* [1] dataset on nested named entity recognition.

3 Conclusions

BioASQ facilitates the exchange and fusion of ideas, providing unique realistic datasets and evaluation services for research on methods for biomedical information access. Therefore, it eventually accelerates progress in the field, as indicated by the gradual improvement of the scores achieved in its tasks [14, 13, 18]. An illustrative example is the contribution of BioASQ in the adoption of fully automated MeSH indexing in NLM [4]. Similarly, we expect that the new version of BioASQ will allow the participating teams to bring further improvement to the six open tasks offered in this edition.

4 Acknowledgments

Google was a sponsor of BioASQ in 2024. This edition of BioASQ was also sponsored by Ovid Technologies, Inc., Elsevier, and Atypon Systems inc. This research was funded by the Spanish National BARITONE project (TED2021-129974B-C22). This work is also supported by the European Union's Horizon Europe Co-ordination & Support Action under Grant Agreement No 101080430 (AI4HF), as well as Grant Agreement No 101057849 (DataTool4Heartproject). The work on the *BioNNE-L* task was supported by the Russian Science Foundation [grant number 23-11-00358]. This work is partially supported by the HEREDITARY Project, as part of the European Union's Horizon Europe research and innovation programme under grant agreement No GA 101137074.

References

1. Davydova, V., Loukachevitch, N., Tutubalina, E.: Overview of bionne task on biomedical nested named entity recognition at bioasq 2024. In: CLEF Working Notes (2024), https://ceur-ws.org/Vol-3740/paper-03.pdf

[13] https://github.com/bioasq
[14] http://participants-area.bioasq.org/datasets

2. Gasco, L., Nentidis, A., Krithara, A., Estrada-Zavala, D., Murasaki, R.T., Primo-Peña, E., Bojo Canales, C., Paliouras, G., Krallinger, M., et al.: Overview of BioASQ 2021-MESINESP track. Evaluation of advance hierarchical classification techniques for scientific literature, patents and clinical trials. CEUR Workshop Proceedings (2021)
3. Krallinger, M., Krithara, A., Nentidis, A., Paliouras, G., Villegas, M.: BioASQ at CLEF2020: Large-scale biomedical semantic indexing and question answering. In: European Conference on Information Retrieval. pp. 550–556. Springer (2020)
4. Krithara, A., Mork, J.G., Nentidis, A., Paliouras, G.: The road from manual to automatic semantic indexing of biomedical literature: a 10 years journey. Frontiers in Research Metrics and Analytics **8** (2023). https://doi.org/10.3389/frma.2023.1250930
5. Krithara, A., Nentidis, A., Bougiatiotis, K., Paliouras, G.: BioASQ-QA: A manually curated corpus for Biomedical Question Answering. bioRxiv (2022)
6. Krithara, A., Nentidis, A., Paliouras, G., Krallinger, M., Miranda, A.: BioASQ at CLEF2021: Large-Scale Biomedical Semantic Indexing and Question Answering. In: European Conference on Information Retrieval. pp. 624–630. Springer (2021)
7. Krithara, A., Nentidis, A., Vandorou, E., Katsimpras, G., Almirantis, Y., Arnal, M., Bunevicius, A., Farre-Maduell, E., Kassiss, M., Konstantakos, V., Matis-Mitchell, S., Polychronopoulos, D., Rodriguez-Pascual, J., Samaras, E.G., Samiotaki, M., Sanoudou, D., Vozi, A., Paliouras, G.: BioASQ Synergy: a dialogue between question-answering systems and biomedical experts for promoting COVID-19 research. Journal of the American Medical Informatics Association p. ocae232 (08 2024). https://doi.org/10.1093/jamia/ocae232, https://doi.org/10.1093/jamia/ocae232
8. Lima-López, S., Farré-Maduell, E., Gascó, L., , Nentidis, A., Krithara, A., Katsimpras, G., Paliouras, G., Krallinger, M.: Overview of MedProcNER Task on Medical Procedure Detection and Entity Linking at BioASQ 2023. In: CEUR Workshop Proceedings (2023)
9. Lima-López, S., Farré-Maduell, E., Rodríguez-Miret, J., Rodríguez-Ortega, M., Lilli, L., Lenkowicz, J., Ceroni, G., Kossoff, J., Shah, A., Nentidis, A., Krithara, A., Katsimpras, G., Paliouras, G., Krallinger, M.: Overview of multicardioner task at bioasq 2024 on medical specialty and language adaptation of clinical ner systems for spanish, english and italian. In: Conference and Labs of the Evaluation Forum (2024), https://ceur-ws.org/Vol-3740/paper-02.pdf
10. Loukachevitch, N., Manandhar, S., Baral, E., Rozhkov, I., Braslavski, P., Ivanov, V., Batura, T., Tutubalina, E.: Nerel-bio: a dataset of biomedical abstracts annotated with nested named entities. Bioinformatics **39**(4), btad161 (2023)
11. Malakasiotis, P., Pavlopoulos, I., Androutsopoulos, I., Nentidis, A.: Evaluation measures for task b. Tech. rep., Tech. rep. BioASQ (2018), http://participants-area.bioasq.org/Tasks/b/eval_meas_2018
12. Miranda-Escalada, A., Gascó, L., Lima-López, S., Farré-Maduell, E., Estrada, D., Nentidis, A., Krithara, A., Katsimpras, G., Paliouras, G., Krallinger, M.: Overview of DisTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings (2022)
13. Nentidis, A., Katsimpras, G., Krithara, A., Lima López, S., Farré-Maduell, E., Gasco, L., Krallinger, M., Paliouras, G.: Overview of BioASQ 2023: The Eleventh BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Giachanou,

A., Li, D., Aliannejadi, M., Vlachos, M., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 227–250. Springer Nature Switzerland, Cham (2023)

14. Nentidis, A., Katsimpras, G., Krithara, A., Lima-López, S., Farré-Maduell, E., Krallinger, M., Loukachevitch, N., Davydova, V., Tutubalina, E., Paliouras, G.: Overview of BioASQ 2024: The twelfth BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. In: Goeuriot, L., Mulhem, P., Quénot, G., Schwab, D., Soulier, L., Maria Di Nunzio, G., Galuščáková, P., García Seco de Herrera, A., Faggioli, G., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024) (2024)
15. Nentidis, A., Katsimpras, G., Krithara, A., Paliouras, G.: Overview of BioASQ Tasks 11b and Synergy11 in CLEF2023. In: CEUR Workshop Proceedings (2023)
16. Nentidis, A., Katsimpras, G., Krithara, A., Paliouras, G.: Overview of bioasq tasks 12b and synergy12 in clef2024. In: Working Notes of CLEF (2024), https://ceur-ws.org/Vol-3740/paper-01.pdf
17. Nentidis, A., Katsimpras, G., Vandorou, E., Krithara, A., Gasco, L., Krallinger, M., Paliouras, G.: Overview of BioASQ 2021: The Ninth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 239–263. Springer (2021)
18. Nentidis, A., Katsimpras, G., Vandorou, E., Krithara, A., Miranda-Escalada, A., Gasco, L., Krallinger, M., Paliouras, G.: Overview of BioASQ 2022: The Tenth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 13390 LNCS, pp. 337–361 (oct 2022). https://doi.org/10.1007/978-3-031-13643-6_22
19. Nentidis, A., Katsimpras, G., Vandorou, E., Krithara, A., Paliouras, G.: Overview of BioASQ Tasks 9a, 9b and Synergy in CLEF2021. In: Proceedings of the 9th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering. CEUR Workshop Proceedings (2021), http://ceur-ws.org/Vol-2936/paper-10.pdf
20. Nentidis, A., Katsimpras, G., Vandorou, E., Krithara, A., Paliouras, G.: Overview of BioASQ Tasks 10a, 10b and Synergy10 in CLEF2022. In: CEUR Workshop Proceedings. vol. 3180, pp. 171–178 (2022)
21. Nentidis, A., Krithara, A., Paliouras, G., Farre-Maduell, E., Lima-Lopez, S., Krallinger, M.: BioASQ at CLEF2023: The Eleventh Edition of the Large-Scale Biomedical Semantic Indexing and Question Answering Challenge. In: Kamps, J., Goeuriot, L., Crestani, F., Maistro, M., Joho, H., Davis, B., Gurrin, C., Kruschwitz, U., Caputo, A. (eds.) Advances in Information Retrieval. pp. 577–584. Springer Nature Switzerland, Cham (2023)
22. Nentidis, A., Krithara, A., Paliouras, G., Gasco, L., Krallinger, M.: BioASQ at CLEF2022: The Tenth Edition of the Large-scale Biomedical Semantic Indexing and Question Answering Challenge. In: European Conference on Information Retrieval. pp. 429–435. Springer (2022)
23. Nentidis, A., Krithara, A., Paliouras, G., Krallinger, M., Sanchez, L.G., Lima, S., Farre, E., Loukachevitch, N., Davydova, V., Tutubalina, E.: BioASQ at CLEF2024: The Twelfth Edition of the Large-Scale Biomedical Semantic Indexing and Question Answering Challenge. In: Advances in Information Re-

trieval. Springer Nature Switzerland, Springer Nature Switzerland, Cham (2024), https://link.springer.com/chapter/10.1007/978-3-031-56069-9_67
24. Ngomo, A.C.N., Heino, N., Speck, R., Ermilov, T., Tsatsaronis, G.: Annotation tool. Project deliverable D3.3 (02/2013 2013), http://www.bioasq.org/sites/default/files/PublicDocuments/2013-D3.3-AnnotationTool.pdf
25. Rodriguez-Penagos, C., Nentidis, A., Gonzalez-Agirre, A., Asensio, A., Armengol-Estapé, J., Krithara, A., Villegas, M., Paliouras, G., Krallinger, M.: Overview of MESINESP8, a Spanish Medical Semantic Indexing Task within BioASQ 2020 (2020)
26. Salton, G., Buckley, C.: Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science **41**(4), 288–297 (jun 1990). https://doi.org/10.1002/(SICI)1097-4571(199006)41:4¡288::AID-ASI8¿3.0.CO;2-H
27. Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M.R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., Almirantis, Y., Pavlopoulos, J., Baskiotis, N., Gallinari, P., Artières, T., Ngonga, A., Heino, N., Gaussier, E., Barrio-Alvers, L., Schroeder, M., Androutsopoulos, I., Paliouras, G.: An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinformatics **16**, 138 (2015). https://doi.org/10.1186/s12859-015-0564-6
28. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)