

RESEARCH NOTE

Open Access



Dialogues of delivery: a multilingual question-answer dataset for maternal healthcare in East African languages

Richard Kimera¹, Rogers Kuyeso², Emmanuel Maleka³, Fortunate Tukamushaba¹, Wasswa William⁴, Rogers Mwavu¹, David Sabiiti Bamutura⁵, Joseph Ngonzi⁶, Brenda Ainomugisha⁶, Alexer Namuli⁶, Proscovia Namiro⁶, Clare Nyakato⁷, Leo Anthony Celi^{8,9,10*} and Fred Kaggwa^{5*}

Abstract

Objective There is a critical scarcity of domain-specific, clinically grounded Natural Language Processing (NLP) resources for African languages. In Western Uganda, linguistic diversity creates a barrier to maternal healthcare, as mothers lack access to health information in their native languages. The objective of this dataset is to provide a high-quality, in-language medical corpus to enable the development and fine-tuning of Large Language Models (LLMs) and conversational AI tools for maternal health in resource-constrained settings.

Data description The “Dialogues of Delivery” dataset is a multilingual, parallel corpus comprising 3,694 question-and-answer pairs presented in four languages: English, Luganda, Runyankore, and Swahili (14,800 total entries). Using facility-based convenience sampling at two health facilities in Western Uganda, primary data was collected via structured, open-ended questionnaires from 150 participants (expectant/postpartum mothers and maternal healthcare providers). The dataset underwent a rigorous forward-backward translation protocol by certified linguists and human-in-the-loop clinical validation by independent medical professionals. The dataset captures core maternal health domains, providing a culturally and clinically validated foundation for Afrocentric AI development.

Keywords Maternal health, Multilingual health communication, Clinical dialogue dataset, Large Language Models, Conversational artificial intelligence, Low-resource settings

*Correspondence:

Leo Anthony Celi

lceli@mit.edu

Fred Kaggwa

kaggwa_fred@must.ac.ug

¹Department of Information Technology, Faculty of Computing and Informatics, Mbarara University of Science and Technology, Mbarara, Uganda

²Department of Software and Informatics Engineering, Faculty of Computing and Informatics, Mbarara University of Science and Technology, Mbarara, Uganda

³Department of Pharmaceutical Sciences, Faculty of Medicine, Mbarara University of Science and Technology, Mbarara, Uganda

⁴Department of Biomedical Engineering, Faculty of Applied Sciences and Technology, Mbarara University of Science and Technology, Mbarara, Uganda

⁵Department of Computer Science, Faculty of Computing and Informatics, Mbarara University of Science and Technology, Mbarara, Uganda

⁶Department of Obstetrics and Gynecology, Faculty of Medicine, Mbarara University of Science and Technology, Mbarara, Uganda

⁷Department of Education, Faculty of Education and Sciences, University of Saint Joseph, Mbarara, Uganda

⁸Massachusetts Institute of Technology, Cambridge, MA, USA

⁹Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA

¹⁰Department of Biostatistics, Harvard T H Chan School of Public Health, Boston, MA, USA



© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Objective

Maternal health outcomes in East Africa remain a pressing concern, with high mortality rates persisting in the region [1, 2]. In rural and peri-urban areas, mothers face numerous barriers to care, among which language and communication gaps are especially critical [3–5]. In linguistically diverse countries like Uganda, which has over 40 languages [6], clinical encounters frequently occur in nonnative languages. This linguistic mismatch compromises diagnostic accuracy, shared decision-making, and patient trust [7–9], ultimately elevating the risk of adverse outcomes and delayed emergency care [10, 11].

While existing mHealth platforms have improved healthcare access [12–14], they often lack interactivity and exclude non-literate populations [15]. Large Language Models (LLMs) offer promising solutions for accessible, conversational patient education. However, their effectiveness in low-resource environments is fundamentally constrained by a lack of high-quality, domain-specific training datasets in local languages [16]. Existing models remain overwhelmingly English-centric. Although multilingual African NLP initiatives are growing [17–19], there remains a severe scarcity of clinically grounded, in-language corpora specifically focused on maternal health [20].

To address this critical gap, we introduce the *Dialogues of Delivery* dataset: a multilingual, clinically validated question-and-answer corpus in English, Luganda, Runyankore, and Swahili. The primary objective of this dataset is to support the fine-tuning of LLMs for prenatal and postnatal care in East African languages. By providing a culturally tailored and medically accurate foundation, this work aims to enable the development of conversational AI tools that are linguistically appropriate. This resource bridges critical data gaps in low-resource settings, contributing to an equitable digital health ecosystem that empowers mothers and frontline healthcare workers with accessible information.

Data description

The dataset captures maternal health dialogues to facilitate conversational AI development in East African languages. Data was gathered at Mbarara Regional Referral Hospital and Kabwohe Health Centre IV in Western Uganda. Participants were recruited via facility-based convenience sampling, yielding 150 individuals: 65 expectant mothers, 45 postpartum mothers, and 40 healthcare workers (clinicians, midwives, nurses, and Village Health Team members). Primary data collection was conducted over a two-week period by six trained

research assistants. To capture rich dialogues, collection utilized multiple formats: semi-structured private interviews ($n = 90$, 60%) averaging 35 min, focus group discussions ($n = 8$ groups) lasting approximately 90 min, and strictly anonymous questionnaires ($n = 52$, 34.7%). Interactive sessions were audio-recorded with consent, transcribed verbatim, and supplemented by observational field notes.

We simulated authentic clinical interactions by having healthcare workers answer questions typically posed by mothers, and vice versa. The questions were systematically categorized into prenatal and postnatal themes. To ensure balanced representation, each thematic category was explicitly structured to contain a minimum of 250 question-answer pairs. The raw responses were digitized, and all Protected Health Information (PHI) was removed to ensure anonymity. A dedicated linguist generated a parallel dataset in English, Luganda, Runyankore, and Swahili using a forward-backward translation protocol (initial translation, independent back-translation, and consensus reconciliation). Translation quality was validated via semantic similarity metrics, achieving average BLEU scores of 0.78 (English-Luganda), 0.81 (English-Runyankore), and 0.83 (English-Swahili).

Subsequently, three independent medical professionals (an obstetrician, a midwife, and a public health specialist) performed human-in-the-loop validation to guarantee clinical integrity and cultural relevance. This audit resulted in the revision of 127 question-answer pairs (3.4%).

The finalized corpus contains 3,694 unique Q/A pairs, totaling 14,800 parallel entries across the four languages. The average question length is 12.4 words, and the average answer length is 38.7 words. The dataset file is approximately 8.2 MB and utilizes UTF-8 encoding to accurately preserve special characters. Thematically, the data highlights the intersection of medical advice and communal life, with the largest segments focusing on community and cultural considerations (31.4%), medical history and lifestyle (24.5%), and symptoms and concerns (24.4%). The high prevalence of cultural inquiries aligns with evidence that mothers navigate complex tensions between traditional practices and biomedical recommendations [21], often driven by deep-rooted cultural trust [22]. Furthermore, the smaller subset of mental health inquiries (6.5%) underscores the clinical reality of stigma, limited mental health literacy, and structural barriers surrounding perinatal depression in the region [23]. An overview of the data file is provided in Table 1.

Table 1 Overview of data files/data set

Label	Name of data file/data set	File types (extension)	Data repository and identifier (DOI or accession number)
Data file 1	Maternal_Multilingual_Dataset_Uganda_2026	.xls	https://doi.org/10.7910/DVN/JS3ILT

Limitations

While this dataset provides a culturally resonant baseline for low-resource languages, it has some limitations. Primarily, the data exhibits limited generalizability, as it is confined to a specific linguistic and geographical scope. Consequently, it may not fully capture the dialectal variations, diverse challenges, or cultural nuances present in other regions of East Africa. Future work should expand upon this high-quality baseline to include broader regional dialects and audio-based modalities. Finally, any conversational AI trained on this dataset must be positioned strictly as a supplementary information-dissemination tool. It is intended to complement, rather than supplant, consultation with qualified healthcare professionals.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13104-026-07899-z>.

Supplementary Material 1.

Acknowledgements

The authors would like to extend their sincere gratitude to the international collaborators who provided invaluable expertise and guidance, including MIT Critical Data, the Data Science for Social Impact Research Group at the University of Pretoria, and Handong Global University (Korea). We are also deeply grateful for the partnership and support of the Mbarara Regional Referral Hospital and Kabwohe Health Centre IV, which served as the performance sites for this research. Finally, our heartfelt thanks go to the expectant mothers, postpartum mothers, Community Health Workers (CHWs), and local leaders whose participation and insights were central to the success of this project.

Author contributions

R.K. contributed to project administration, conceptualization, supervision, and writing of the original draft. F.K. contributed to project administration, methodology, data curation, investigation, and resources. R.Kuyeso contributed to software development, methodology, and investigation. E.M. contributed to investigation, and resources. D.S.B. contributed to data curation, investigation, and validation. R.M. and W.W. contributed to formal analysis and visualization. F.T. contributed to project administration and resources. J.N. & B.A. contributed domain-specific medical expertise, including the design and validation of data collection instruments and survey frameworks. A.N. & P.N. were responsible for data collection, clinical validation, and comprehensive review of the dataset, with particular involvement in midwifery-related components. C.N. contributed to the translation of the dataset and its linguistic validation to ensure adherence to standardized and contextually appropriate language. L.A.C. contributed to conceptualization, methodology, and supervision. All authors reviewed and approved the final manuscript.

Funding

This work was supported by the Science for Africa (SFA) Foundation through the Grand Challenges Africa program under Grant Reference Number SFA-R15-145.

Data availability

The primary dataset developed in this study has been deposited in the Harvard Dataverse Repository and is accessible at the following link: <https://doi.org/10.7910/DVN/JS3ILT>.

Declarations

Ethics approval and consent to participate

Institutional approval was obtained from Mbarara University of Science and Technology Research Ethics Committee (MUST-REC) (Reference Number: MUST-2024-1779) and national clearance from the Uganda National Council for Science and Technology (UNCST) (Reference Number: HS5739ES). All procedures involving human participants were conducted in accordance with the principles of the Declaration of Helsinki. Written informed consent was obtained from all participants prior to data collection.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 6 February 2026 / Accepted: 26 May 2026

Published online: 02 June 2026

References

1. Yihune Teshale M, Bante A, Gedefaw Belete A, Crutzen R, Spigt M, Stutterheim SE. Barriers and facilitators to maternal healthcare in East Africa: a systematic review and qualitative synthesis of perspectives from women, their families, healthcare providers, and key stakeholders. *BMC Pregnancy Childbirth*. 2025;25(1):111.
2. Birabwa C, Banke-Thomas A, Waiswa P, Semaan A, Kananura RM, van Olmen J, Beňová L. Maternal health in cities: analysis of institutional maternal mortality and health system bottlenecks in Kampala City Uganda, 2016–2021. *J Global Health Rep*. 2024;8.
3. Ssetaala A, Ssempiira J, Nanyonjo G, Okech B, Chinyenze K, Bagaya B, Price MA, Kiwanuka N, Degomme O. Mobility for maternal health among women in hard-to-reach fishing communities on Lake Victoria, Uganda; a community-based cross-sectional survey. *BMC Health Serv Res*. 2021;21(1):948.
4. Kofo KW, Matiang'i M, Akolo M. Barriers to utilization of skilled birth attendance among post-natal women in a North Horr ward, Marsabit County, Kenya: A cross-sectional study. *Edorium J Matern Child Health*. 2024;8(2):1–1.
5. Kimera R, Heo D, Rim DN, Choi H. Data Augmentation With Back translation for Low Resource languages: A case of English and Luganda. In: *Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval*. 2024;142–148.
6. Ndayizigamiye P. Implementation of mHealth Interventions for Public Healthcare in East Africa: Lessons Learnt. In: *Building Resilient Healthcare Systems With ICTs*. IGI Global Scientific Publishing. 2022;106–24.
7. Al Shamsi H, Almutairi AG, Al Mashrafi S, Al Kalbani T. Implications of language barriers for healthcare: a systematic review. *Oman Med J*. 2020;35(2):e122.
8. Knop MR, Nagashima-Hayashi M, Lin R, Saing CH, Ung M, Oy S, Yam EL, Zahari M, Yi S. Impact of mHealth interventions on maternal, newborn, and child health from conception to 24 months postpartum in low-and middle-income countries: a systematic review. *BMC Med*. 2024;22(1):196.
9. Nakijoba R, Twimukye A, Bayigga J, Kawuma AN, Asimwe SP, Byenume F, Kyeyune J, Nabukenya S, Ojara FW, Waitt C. Culturally appropriate terminologies in health research: a participatory study with pregnant and breastfeeding mothers in Uganda. *Front Communication*. 2025;9:1450569.
10. Twimukye A, Nabukenya S, Kawuma AN, Bayigga J, Nakijoba R, Asimwe SP, Byenume F, Ojara FW, Waitt C. Some parts of the consent form are written using complex scientific language: community perspectives on informed consent for research with pregnant and lactating mothers in Uganda. *BMC Med Ethics*. 2024;25(1):149.
11. Macharia P, Pérez-Navarro A, Sambai B, Inwani I, Kinuthia J, Nduati R, Carrión C. An unstructured supplementary service data-based mhealth app providing on-demand sexual reproductive health information for adolescents in kibra, kenya: randomized controlled trial. *JMIR mHealth uHealth*. 2022;10(4):e31233.
12. Hailemariam T, Atnafu A, Gezie LD, Tilahun B. Effect of short message service reminders in improving optimal antenatal care, skilled birth attendance and postnatal care in low-and middle-income countries: a systematic review and meta-analysis. *BMC Med Inf Decis Mak*. 2024;25(1):1.

13. Mekonnen ZA, Gelaye KA, Were M, Tilahun B. Effect of mobile phone text message reminders on the completion and timely receipt of routine childhood vaccinations: superiority randomized controlled trial in Northwest Ethiopia. *JMIR mHealth uHealth*. 2021;9(6):e27603.
14. Musiimenta A, Tumuhimbise W, Atukunda EC, Mugaba AT, Asasira J, Katusiime J, Zender R, Pinkwart N, Mugenyi GR, Haberer JE. A mobile health app may improve maternal and child health knowledge and practices among rural women with limited education in Uganda: a pilot randomized controlled trial. *JAMIA open*. 2022;5(4).
15. Richter-Pechanski P, Wiesenbach P, Schwab DM, Kiriakou C, Geis N, Dieterich C, Frank A. Clinical information extraction for lower-resource languages and domains with few-shot learning using pretrained language models and prompting. *Nat Lang Process* 2024;31:1–24.
16. Orife I, Kreutzer J, Sibanda B, Whitenack D, Siminyu K, Martinus L, Ali JT, Abbott J, Marivate V, Kabongo S, Meressa M. Masakhane—machine translation for Africa. *arXiv preprint arXiv:2003.11529*. 2020.
17. Olatunji T, Afonja T, Yadavalli A, Emezue CC, Singh S, Dossou BF, Osuchukwu J, Osei S, Tonja AL, Etori N, Mbataku C. Afrispeech-200: Pan-african accented speech dataset for clinical and general domain asr. *Trans Association Comput Linguistics*. 2023;11:1669–85.
18. Chinyamakobvu M, Govender J, James U, Leslie T, Letuka T, Ncube A. Lacuna Fund learning and evaluation report. Data Innovators & Meridian Institute. 2023. https://lacunafund.org/wp-content/uploads/sites/11/2023/12/Lacunafund_DI_ELReport_FA_Digital-Version_22.09.23.pdf
19. Hu S, Oppong A, Mogo E, Collins C, Occhini G, Barford A, Korhonen A. Natural language processing technologies for public health in Africa: scoping review. *J Med Internet Res*. 2025;27:e68720.
20. Pell C, Menaca A, Were F, Afrah NA, Chatio S, Manda-Taylor L, Hamel MJ, Hodgson A, Tagbor H, Kalilani L, Ouma P. Factors affecting antenatal care attendance: results from qualitative studies in Ghana, Kenya and Malawi. *PLoS ONE*. 2013;8(1):e53747.
21. Shewamene Z, Dune T, Smith CA. The use of traditional medicine in maternity care among African women in Africa and the diaspora: a systematic review. *BMC Complement Altern Med*. 2017;17(1):382.
22. Atuhaire C, Rukundo GZ, Brennaman L, Cumber SN, Nambozi G. Lived experiences of Ugandan women who had recovered from a clinical diagnosis of postpartum depression: a phenomenological study. *BMC Pregnancy Childbirth*. 2021;21(1):826.
23. Abajobir A, Sidze EM, Wainaina C, Gerbaba MJ, Wekesah FM, Abajobir, et al. *Arch Women Ment Health*. 2025;28(5):997–1089.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.