Evaluation of an NLP-Based Chatbot for Informational Support in Bronchopulmonary Dysplasia (BPD) in neonates

Anna Beatriz Silva[1],
Cleyton Mario de Oliveira Rodrigues [1], **Patricia Takako Endo**[1]

[1]Programa de Pós-Graduação em Engenharia da Computação (PPGEC),
Universidade de Pernambuco (UPE), Caruaru – PE – Brazil

{anna.beatrizs, cleyton.rodrigues, patricia.endo}@upe.br

Abstract. *Bronchopulmonary Dysplasia (BPD) is a chronic lung condition that affects neonates, especially those born prematurely. Despite its relevance, the lack of accessible and up-to-date information hampers the proper management of this condition for both families and healthcare professionals. This study proposes the development of a chatbot based on Natural Language Processing (NLP) technique, named Soprinho IA, to provide informational and empathetic support to both mothers and healthcare professionals, focusing on BPD information. The methodology involves data collection from various sources, text preprocessing, and training of three different NLP architectures, such as Recurrent Neural Networks (RNNs), transformers, and other state-of-the-art approaches. The Sentence-Transformer model, based on BERT, achieved the best results, with high accuracy and semantic relevance. The chatbot is integrated into a responsive web interface, offering multimodal interaction features, including audio support. The evaluation of the models demonstrated the chatbot's effectiveness in providing contextualized and easily understandable responses, with potential to improve communication and the management of neonates with BPD. The study also highlights limitations, such as the need to improve speech-to-text conversion, and suggests future directions for the integration of AI models to support clinical decision-making.*

1. Introduction

Bronchopulmonary dysplasia (BPD) is a chronic lung condition affecting neonates, particularly those born prematurely at less than 37 weeks of gestation [(WHO) 2023]. The World Health Organization (WHO) estimates that approximately 15 million babies are born prematurely each year, with about 10% of these babies classified as very preterm (less than 32 weeks) and 5% as extremely preterm (less than 28 weeks). The global incidence of BPD among preterm newborns ranges from 10% to 89%, according to clinical and observational studies [Siffel et al. 2021, Stoll et al. 2010, Bancalari et al. 2003]. Factors such as mechanical ventilation and prolonged oxygen therapy, respiratory infections, surfactant dysfunction, low birth weight, and pulmonary or cardiac anomalies also contribute to the development of BPD [Thébaud et al. 2019].

Despite its high prevalence, BPD remains a challenge for both healthcare professionals and the families of affected patients. Most mothers of babies affected by the condition often lack the necessary knowledge to manage the complications associated

with this chronic disease. The lack of accessible and tools to disseminate information about BPD exacerbates this situation.

Although there are institutional resources that provide general information about the disease, such as the American Thoracic Society manual, which addresses parental care and professional support after discharge [American Thoracic Society 2013], and specialized programs at referral centers [Stanford Children's Health], these materials are mostly static, such as PDFs and informational manuals. They do not offer an interactive channel to answer specific questions from mothers in real-time, nor empathetic support during the vulnerable post-discharge period.

In recent years, perinatal literature has explored the use of chatbots to support maternal and infant health. However, no existing tools are specifically designed to address the unique and ongoing demands of BPD [Amil et al. 2025]. Given the complexity and need for tailored information, there is a clear gap for an intelligent, dynamic solution in this space.

To address this gap, we propose the development of an interactive chatbot, named Soprinho IA, designed specifically for BPD information, capable of communicating in two distinct modes: one for mothers, offering accessible, empathetic, and easy-to-understand information; and another for healthcare professionals, providing more technical, evidence-based content. This dual-channel approach ensures that both user groups receive the type and depth of information appropriate to their needs and expertise. The proposed chatbot will be developed using traditional Natural Language Processing (NLP) techniques, allowing users to interact simply by asking questions and receiving clear and consistent answers. The goal is to provide informational support during the most challenging times for families and offer an accessible source of technical information for healthcare professionals regarding causes, treatments, and recommended approaches in the follow-up of neonates with BPD.

2. Related Works

Although BPD has not been addressed in chatbots developed so far, the literature related to the use of conversational agents in the perinatal area has shown promising advancements. A recent systematic review identified 12 initiatives of conversational agents targeting the period from preconception to 12 months postpartum. These studies indicated a high level of satisfaction and improvements in health behavior among mothers [Amil et al. 2025].

Rivera et al. (2024) developed a chatbot aimed at postpartum women and caregivers of newborns, observing good acceptance among users, particularly in managing anxiety and identifying early clinical signs [Rivera Rivera et al. 2024]. Another example is the "Rosie" chatbot, which provides personalized responses about pregnancy, maternity, and infant care based on reliable sources such as the CDC and the Mayo Clinic, demonstrating high engagement among users [Nguyen et al. 2024]. Leitner et al. (2025) presented an NLP-based agent for the postpartum period, which achieved 98% interaction among mothers and provided informational support on breastfeeding, physical recovery, and risk signs for neonates [Leitner et al. 2025].

These studies highlight the potential of chatbots to offer informational and emotional support to mothers and caregivers in the perinatal context. Their evaluation has

been based on a variety of metrics, including the rate of user engagement through message opening and interaction, levels of satisfaction and usability measured by standardized surveys, and the perceived usefulness of reminders and educational content. In addition, some studies analyzed behavioral outcomes such as adherence to medical appointments and the impact on health practices. Qualitative data, obtained through structured interviews, also contributed to assessing the relevance and acceptance of the tools. However, none of these chatbots are targeted at bronchopulmonary dysplasia. This condition requires continuous monitoring and specific technical information. Therefore, the aim of this study is to fill this gap by offering an interactive solution specialized in BPD.

3. Materials and Methods

The development of the chatbot followed a meticulous approach based on NLP models, aimed at creating an effective tool that caters to both the general public, such as the mothers being attended, and healthcare professionals.

An audio functionality was integrated into the system after the models implementation. Users can send audio, which is transcribed into text. This transcribed text is then processed by the NLP model to generate an appropriate response. Furthermore, the response can be converted back into audio, enabling a multimodal interaction. This feature facilitates communication, especially for users with reading difficulties, and aims to enhance accessibility while optimizing the overall user experience.

The methodology was structured into seven main stages, as detailed in Figure 1.

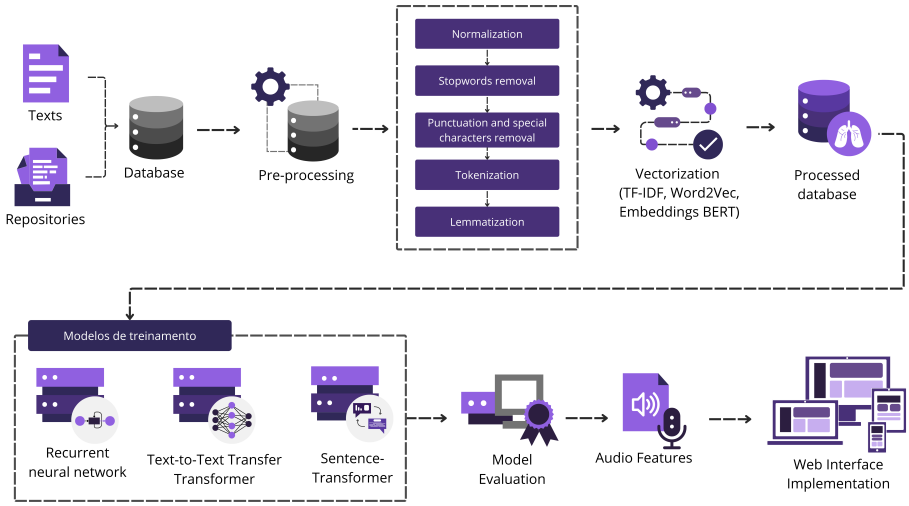

Figure 1. Steps of the chatbot development methodology

3.1. Data sources and pre-processing

The first step involved collecting the data used, sourced from various platforms, including institutional repositories, scientific articles, and educational materials. These documents formed the initial textual base, which was then organized into a database to be processed and analyzed.

The first step involved collecting the data used, sourced from various platforms, including institutional repositories, scientific works, and educational materials. These

documents formed the initial textual base, which was then organized into a database to be processed and analyzed. Two primary sources were: the MSD Manual, Professional Version [Balest 2023a], which provides technical and detailed information aimed at healthcare professionals, and the MSD Manual, Patient Version [Balest 2023b], which presents the same core content in a simplified and more accessible language for the general public. These two versions allowed for a comparative textual base, supporting the design of differentiated responses for distinct user profiles, a process aligned with principles of textual complexity and audience adequacy discussed before [Leal and Aluísio 2024].

Given the limited availability of literature in Portuguese, additional materials were drawn from international scientific articles on neonatal care and BPD management. Selected information from these works was translated and adapted into Portuguese, ensuring consistency in terminology, preserving technical accuracy, and making the content comprehensible to non-specialist readers.

The data collection process was conducted manually through a review of the selected sources, from which the information was extracted for use in model training. The extracted content was organized into CSV files, forming the initial dataset. Each record contained the following fields: informational content, user profile, and thematic category. A manual categorization step was also carried out, assigning metadata to each entry to indicate its tone, intended audience, and subject relevance. This structured approach enabled the models to differentiate responses based on user type and contextual needs.

Following data collection and organization, the preprocessing phase was carried out, including steps such as text normalization, removal of stopwords, punctuation, and special characters, as well as tokenization and lemmatization, aiming to reduce words to their base forms. Subsequently, text vectorization techniques, such as Word2Vec, TF-IDF, and embeddings based on BERT-like models, were applied to convert the textual data into numerical representations suitable for training the models.

3.2. NPL models

Three distinct NLP models were employed, each with a specific approach to generating responses to predefined questions. The first model used was T5 (Text-to-Text Transfer Transformer), which uses the Transformer architecture [Raffel et al. 2020]. This model treats all NLP tasks as text transformation problems, where the input is a question and the output is a response generated from that input.

The second model was the Recurrent Neural Network (RNN), a technique designed to handle sequential data, such as text [Lipton et al. 2015]. RNNs process words sequentially, retaining previous information to generate the response. The third model was the Sentence-Transformer, based on BERT variants, which generates vector embeddings for questions and answers [Reimers and Gurevych 2019]. This model transforms the inputs into high-dimensional vectors, allowing semantic similarity between them to be measured using cosine similarity.

Although the data and models used are the same for both audiences, in a future version of the chatbot, doctors will be able to access more technical and detailed responses about BPD. At the same time, the basic interaction structure and databases will remain consistent for both users.

3.3. Evaluation metrics

Initially, the training process for these models involved dividing the dataset into training, validation, and test sets. The training set, which accounted for 70% of the data, was used to teach the models how to generate answers. The validation set represents 10% of the data, responsible for helping tune hyperparameters and avoid overfitting. The test set, which was kept completely separate and made up the remaining 20%, was used to evaluate the final performance of the models.

The performance of the models was evaluated based on quantitative metrics such as accuracy, precision, recall, and F1-score. This allowed for the model selection with the best predictive performance and the highest generalization capacity in simulated textual interactions.

Once the best-performing model was identified, the model was then integrated into a responsive web interface, accessible from various devices, ensuring an inclusive user experience. The interface was designed to dynamically adapt responses according to the interlocutor's profile, ensuring that information is delivered appropriately for the general public and healthcare professionals. The interface offers two versions: one more technical for doctors and another more simplified for mothers, ensuring clear and efficient communication for each audience.

The dataset is structured with three main fields. The *question_text* field contains the question asked by the user, while the *response_text* field generates the most appropriate answer generated by the system. The profile field indicates the intended audience for that response. This profile can be labeled as either professional or user. Users accessing the chatbot as patients or caregivers do not need to log in; in these cases, the chatbot automatically adopts a simplified, empathetic tone tailored to a general audience. In contrast, healthcare professionals must log in to access content with a more technical depth. Once logged in, the chatbot provides responses using specialized language and clinical evidence, tailored to the needs of professional healthcare practice.

4. Results

4.1. Models performance

The T5 model demonstrated a good semantic similarity with an average of 0.8289, indicating that the generated answers were semantically relevant. However, the model struggled to generate exact responses. The accuracy was 76.28%, suggesting that the model's performance was reasonable, although it struggled to match the correct answers in some cases. Precision and recall were also low, with an F1-Score of 0.43, indicating that the model had difficulties generating fully accurate responses, despite producing semantically good ones.

In comparison, the RNN model achieved an accuracy of 78.21% and an F1-Score of 0.61. The model still struggled to capture the deep semantic context of the responses, resulting in suboptimal performance for tasks that require a better understanding of semantics.

By contrast, the Sentence-Transformer was the best-performing model, achieving an accuracy of 93.59% and an F1-Score of 0.9289. The model also had a mean cosine similarity of 0.8289, suggesting that the generated responses were very close to the actual

answers in terms of meaning. Precision and recall were also higher in the Sentence-Transformer, indicating that the model was well-balanced in identifying the correct responses.

In terms of semantic similarity, both the T5 and Sentence-Transformer models showed a good ability to generate responses that were semantically coherent with the actual answers. The results of these evaluations can be observed in Table 1.

Model	Accuracy	Precision	Recall	F1-Score
T5 (Transformer)	0.7628	0.5000	0.3814	0.4327
RNN (SimpleRNN)	0.7821	0.6167	0.6150	0.6121
Sentence-Transformer (BERT)	0.9359	0.9394	0.9576	0.9289

Table 1. Models Performance Results

Based on the evaluation metrics, the Sentence-Transformer model was selected for integration into the chatbot system. Its high accuracy, F1-Score, and balance between precision and recall, combined with the semantic similarity score, indicated that it was the most effective model for generating contextually appropriate and accurate responses.

To evaluate the semantic performance of the selected model, the text generation metrics ROUGE-L F1 [Nenkova 2005] and METEOR [Banerjee and Lavie 2005] were used to measure the quality of the generated responses, focusing on semantic similarity and linguistic quality.

ROUGE-L F1 measures the overlap of the longest subsequences between the generated response and the reference response, while METEOR evaluates semantic match, considering synonyms, word order, and other linguistic aspects.

The Sentence-BERT model achieved a ROUGE-L F1 score of 0.68, indicating that the generated responses match the ideal responses, but there is still room for improvement. The METEOR metric, on the other hand, achieved a performance of 0.633, suggesting that the model generates responses that are relatively good semantically but have not yet reached the ideal level of match in terms of synonyms or linguistic fluency.

The ROUGE-L F1 and METEOR metrics offer valuable insights into the semantic quality of the generated responses. Although the results demonstrate that the model achieves a reasonable overlap and semantic alignment with the reference, there is still potential for improvement. The subsequence overlap could be further optimized through fine-tuning the text generation process, while enhancing semantic coherence may require more extensive training or the inclusion of a more diverse and linguistically varied training dataset.

The audio model was evaluated using the Word Error Rate (WER) and Character Error Rate (CER) metrics, which are commonly employed to assess the accuracy of automatic transcription systems. The WER was 0.3, indicating that 30% of the words in the transcription differed from the reference text. In contrast, the CER was 0.031, suggesting that despite word-level inaccuracies, the model demonstrated high precision in transcribing individual characters. These results imply that the overall transcription was largely accurate.

4.2. Web interface

The responsive web interface developed for the chatbot incorporates functionalities aimed at enhancing user usability and accessibility across diverse user profiles, initially for healthcare professionals and mothers of neonates with the condition. These functionalities were designed following human-centered design principles applied to healthcare [Zhou et al. 2019]. The login screen is shown in Figure 2.

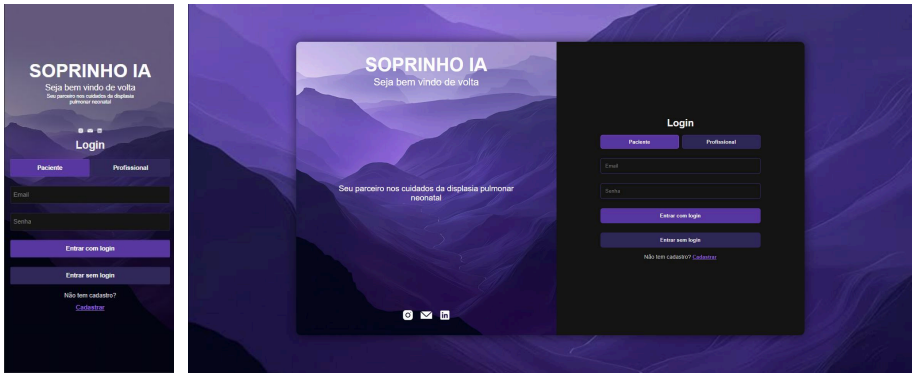

Figure 2. Login screen for desktop with responsive design and elements

Users can start a new interaction session with the chatbot at any time, resetting the conversation context, and allowing the system to generate responses independently of previous interactions. At the same time, the chatbot also stores previous interactions, enabling users to revisit past conversations, and contributing to knowledge retention. Users are also allowed to delete individual conversations, including this privacy functionality.

The chatbot supports voice input, allowing users to send audio messages with a low limit of 40 seconds. These audio inputs are transcribed in real-time processing to text to be underlying by the Sentence-Transformer model. The aim of this functionality is improves accessibility, especially for users with limited education or technological skills. The chat screen is shown in Figure 3.

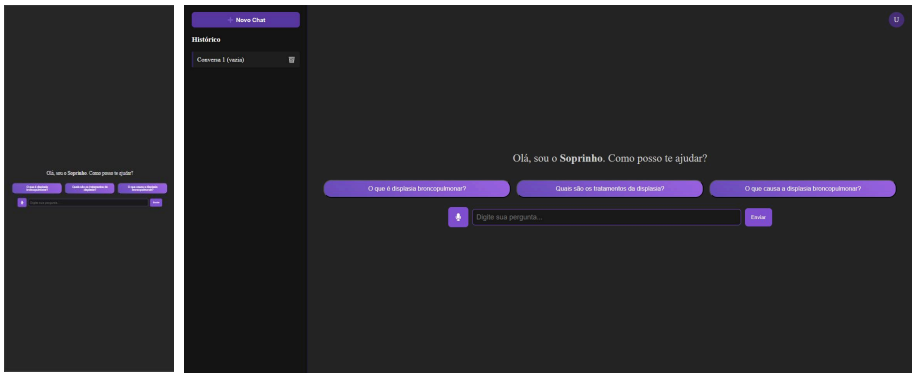

Figure 3. Chatbot screen for desktop with responsive design

Upon login, the system identifies the user as either a healthcare or a mother. Based on this selection, the chatbot adapts the language and depth of information. Medical users receive technically delayed responses, while mothers receive simplified and supportive explanations to understand the condition.

Each of these funcionalities is embedded within a responsivel deisgn structure, sujpport varying screens size. The system architecture includes a RESTful API to ensure real-time interactions and consistent performances.

5. Discussion

The results showed that the Sentence-Transformer model based on BERT outperformed the other models, T5 and RNN, in key metrics. With an accuracy of 93.59% and an F1-Score of 92,89%, the model was effective at generating semantically accurate responses. The high precision of 93.94% and recall of 95.76% further demonstrated that the Sentence-Transformer was particularly good at identifying the correct answers, reducing both false positives and false negatives.

In contrast, the T5 model, while generating semantically relevant answers with a cosine similarity of 0.8289, struggled with accuracy at 76.28% and F1-Score at 43%. These low values indicate that T5 was not as reliable in producing exact answers. Similarly, the RNN model performed better than T5 but still showed limitations. It achieved an accuracy of 78.21% and an F1-Score of 0.6121, suggesting that it struggled with understanding the deeper semantic context needed to generate precise responses.

6. Work Limitations

One of the challenges faced during the development was ensuring the chatbot could handle the nuanced and highly technical language related to BPD. The nature of BPD as a complex medical condition, with varying severity levels and different treatment protocols, meant that the chatbot needed to provide both simplified explanations for the general public and highly technical, evidence-based information for healthcare professionals. While the Sentence-Transformer model performed admirably, future iterations could include more fine-tuned models or additional training to specialize further in medical jargon and more precise medical terms.

Another challenge was the multimodal integration of audio functionality. While the feature of allowing users to interact via voice and receive both text and audio responses was successfully implemented, the accuracy of speech-to-text conversion still requires improvement, particularly in noisy environments. This is important to ensure accessibility for users with reading difficulties and to enhance communication in situations where typing may not be feasible.

7. Conclusion and Future Works

The development and evaluation of the chatbot underscores its potential as a tool for facilitating the exchange of information for mothers and healthcare professionals within the context of BPD. The model's performance demonstrates that NLP-based chatbots can deliver precise and contextually relevant responses in medical settings, thereby enhancing the accessibility and accuracy of medical information.

Looking forward, several areas for further refinement and development can be identified. First, improving the chatbot's capacity to process more complex and open-ended queries would enhance its adaptability and practical utility. While the current version handles predefined questions, the next iteration should focus on enabling the model

to generalize its responses, allowing it to engage with a broader range of inquiries while maintaining contextual relevance.

An additional avenue for advancement involves the integration of supplementary AI models designed to enhance clinical decision-making processes. By incorporating machine learning techniques capable of analyzing patient data, the chatbot could function as a decision-support tool, assisting healthcare professionals by suggesting potential diagnoses, treatment protocols, and follow-up measures based on real-time clinical data.

Moreover, validating the chatbot within real-world clinical environments is preference. Although the current version shows favorable results in controlled settings, its real-world applicability must be assessed through validation with healthcare professionals. Clinician feedback will be instrumental in refining the chatbot's functionality, ensuring that it aligns with clinical workflows and attaches to the standards of medical practice, providing evidence-based, medically accurate recommendations.

The current chatbot represents the first functional version of the application, developed as a proof of concept to demonstrate its feasibility and potential impact in the context of BPD. While it already delivers accurate and contextually relevant responses within predefined scenarios, further adjustments, expanded training datasets, and rigorous validations are necessary before public release

In summary, this study highlights the potential of NLP techniques in developing a chatbot designed to support both patients and healthcare providers in managing BPD. By offering both general and specialized information, the chatbot has the potential to improve communication, information, facilitate enhanced patient care, and contribute to more informed clinical decision-making.

References

American Thoracic Society (2013). *What is Bronchopulmonary Dysplasia (BPD)?* ATS Patient Education Series, online version updated January 2018. Last accessed: June 2025.

Amil, S., Da, S.-M.-A.-R., Plaisimond, J., Roch, G., Sasseville, M., Bergeron, F., and Gagnon, M.-P. (2025). Interactive conversational agents for perinatal health: A mixed methods systematic review. *Healthcare*, 13(4).

Balest, A. L. (2023a). Displasia broncopulmonar (dbp) — versão para profissionais de saúde. Last Accessed: July 2025.

Balest, A. L. (2023b). Displasia broncopulmonar (dbp) — versão para profissionais o paciente. Last Accessed: July 2025.

Bancalari, E., Claure, N., and Sosenko, I. R. (2003). Bronchopulmonary dysplasia: changes in pathogenesis, epidemiology and definition. In *Seminars in neonatology*, volume 8, pages 63–71. Elsevier.

Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Leal, S. E. and Aluísio, S. M. (2024). Complexidade textual e suas tarefas relacionadas. *Processamento de linguagem natural: conceitos, técnicas e aplicações em português*.

Leitner, K., Cutri-French, C., Mandel, A., Christ, L., Koelper, N., McCabe, M., Seltzer, E., Scalise, L., Colbert, J. A., Dokras, A., Rosin, R., and Levine, L. (2025). A conversational agent using natural language processing for postpartum care for new mothers: Development and engagement analysis. *JMIR AI*, 4:e58454.

Lipton, Z. C., Berkowitz, J., and Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*.

Nenkova, A. (2005). Automatic text summarization of newswire: Lessons learned from the document understanding conference. In *AAAI*, volume 5, pages 1436–1441.

Nguyen, Q. C., Aparicio, E. M., Jasczynski, M., Channell Doig, A., Yue, X., Mane, H., Srikanth, N., Gutierrez, F. X. M., Delcid, N., He, X., and Boyd-Graber, J. (2024). Rosie, a health education question-and-answer chatbot for new mothers: Randomized pilot study. *JMIR Form Res*, 8:e51361.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Rivera Rivera, J. N., AuBuchon, K. E., Smith, M., Starling, C., Ganacias, K. G., Danielson, A., Patchen, L., Rethy, J. A., Blumenthal, H. J., Thomas, A. D., and Arem, H. (2024). Development and refinement of a chatbot for birthing individuals and newborn caregivers: Mixed methods study. *JMIR Pediatr Parent*, 7:e56807.

Siffel, C., Kistler, K. D., Lewis, J. F. M., and and, S. P. S. (2021). Global incidence of bronchopulmonary dysplasia among extremely preterm infants: a systematic literature review. *The Journal of Maternal-Fetal & Neonatal Medicine*, 34(11):1721–1731. PMID: 31397199.

Stanford Children's Health. Bronchopulmonary dysplasia program. `https://www.stanfordchildrens.org/en/services/bronchopulmonary-dysplasia`. Last accessed: June 2025.

Stoll, B. J., Hansen, N. I., Bell, E. F., Shankaran, S., Laptook, A. R., Walsh, M. C., Hale, E. C., Newman, N. S., Schibler, K., Carlo, W. A., et al. (2010). Neonatal outcomes of extremely preterm infants from the nichd neonatal research network. *Pediatrics*, 126(3):443–456.

Thébaud, B., Goss, K. N., Laughon, M., Whitsett, J. A., Abman, S. H., Steinhorn, R. H., Aschner, J. L., Davis, P. G., McGrath-Morrow, S. A., Soll, R. F., et al. (2019). Bronchopulmonary dysplasia. *Nature reviews Disease primers*, 5(1):78.

(WHO), W. H. O. (2023). Born too soon: decade of action on preterm birth. Available at: https://www.who.int/publications/i/item/9789240073890.

Zhou, L., Bao, J., Setiawan, I. M. A., Saptono, A., Parmanto, B., et al. (2019). The mhealth app usability questionnaire (mauq): development and validation study. *JMIR mHealth and uHealth*, 7(4):e11500.