## Comments to reviewer 1:

Many thanks for allowing me to read this paper. This paper is of good clinical value and I feel will benefit readers at various levels. I have some comments which you and authors may find useful. The authors present a thorough review of pre- and post-operative voice quality assessment in laryngeal leukoplakia, incorporating subjective, perceptual, objective, and videoendoscopic measures.

1. While it is well established that appropriately performed phonosurgical interventions improve voice outcomes, it is unclear what novel contribution this study adds to the existing literature. But this is still an interesting read.

We thank the reviewer for this valuable comment. While it is well established that phonosurgical interventions can improve voice outcomes, the novelty of our study lies in its specific approach. First, we provide a systematic multidimensional assessment of pre- and postoperative voice function in patients with laryngeal leukoplakia and dysplasia, a group in which functional outcomes are still underrepresented. Second, we apply the Vocal Extent Measure (VEM) longitudinally in this clinical context. To our knowledge, this is the first study to evaluate VEM in these patients and to demonstrate its relationship with both subjective and established objective voice parameters. Third, we explore associations between clinical factors and vocal outcomes, offering additional insight into determinants of postoperative voice function. E.g., the finding that higher-grade dysplasia may be associated with greater postoperative improvement in vocal capacity highlights a potentially relevant clinical pattern. Taken together, the novelty of our work lies not in demonstrating that surgery improves voice per se, but in providing a structured, multidimensional, and quantitatively enhanced evaluation framework, including VEM, and applying it systematically to this specific and clinically relevant patient population.

The authors demonstrate postoperative improvements in voice quality within a heterogeneous cohort, which supports the effectiveness of their surgical approach. VEM has been included as one of the objective assessment tools; this method was originally introduced as a novel technique by the authors in 2018.

2. At least twelve of the cited references are previous publications by the same group, reflecting both their substantial experience in this field and a notable degree of self-citation.

We thank the reviewer for raising this important point. We agree that excessive self-citation should be avoided; however, in this case it reflects the current state of the literature rather than preferential citation. The VEM is a relatively recent parameter, originally introduced by our group, and to date the available body of evidence remains limited. Consequently, the majority of methodological and validation studies on VEM originate from the same research group that developed and systematically evaluated it. Importantly, these citations are directly

relevant, as the present study specifically investigates VEM as a central outcome parameter. At the same time, self-citations account for a minority of the total references (12 out of 77), and were included only where essential to ensure appropriate scientific context and methodological transparency. We have taken care to balance these with the broader literature on voice diagnostics and phonosurgical outcomes. We would also like to note that VEM has been incorporated into national clinical practice recommendations (e.g., AWMF guidelines on voice disorders in Germany), supporting its clinical relevance, even though international adoption is still evolving.

3. Voice quality assessment practices vary considerably across centres. As this paper will be read internationally, it would be helpful to clarify the rationale for selecting VHI-9 instead of VHI-10 or 30, and GRB rather than GRBAS. The discussion explains the use of VEM, but similar justification for the other tools would benefit readers at all levels.

We thank the reviewer for this helpful comment and agree that clarification of the selected assessment tools is important for an international readership. The VHI-9i was developed through item reduction of the original VHI, based on European language versions (Nawka et al., 2009). It was selected for the present study due to its strong psychometric properties and proven clinical utility. In particular, excellent test–retest reliability (r = 0.919) and robust validity in a large cohort with multidimensional voice diagnostics have been demonstrated (Caffier et al., 2021). Importantly, the VHI-9i provides a substantial reduction in administration time for both patients and clinicians compared to the original VHI-30 and other short-form versions (e.g., VHI-12, VHI-10), while retaining comparable clinical information and discriminatory power for grading voice handicap severity. This makes it especially suitable for routine clinical use as well as retrospective cohort analyses such as the present study. Given these advantages, the VHI-9i represents a validated, efficient, and clinically established alternative to longer VHI versions, without compromising the quality of patient-reported outcome assessment.We have added a brief justification for the use of the VHI-9i in the Methods section (2.3. Examiners' Tools and Standards, 2nd para, page 6, line 233 ff) as follows:

“Owing to its validated reliability and validity, reduced administration time, and comparable discriminatory power relative to the original VHI-30 and other longer short-form versions (e.g., VHI-10, VHI-12), the VHI-9i represents an efficient and clinically robust instrument for assessing patient-reported voice handicap.”

The GRB scale was selected as it represents a simplified and widely used perceptual evaluation method, focusing on the core dimensions of overall grade (G), roughness (R), and breathiness (B). This approach is closely related to the RBH classification established by Wendler, which is commonly applied in German-speaking clinical practice and recommended within European frameworks (e.g., ELS protocols). Importantly, the reduction from GRBAS to GRB is not only motivated by practicality but also by methodological considerations. Previous studies have demonstrated that the additional parameters “asthenia” and “strain” show limited reliability and overlap with other dimensions, particularly breathiness, thereby questioning their

independence and clinical utility (e.g., Nawka et al., 1994; Dejonckere et al., 1998; Webb et al., 2004). In contrast, roughness and breathiness reflect the primary pathophysiological components of dysphonia (irregular vocal fold vibration and incomplete glottal closure), while the overall grade captures global voice impairment. Given these considerations, the GRB scale represents a clinically robust, reliable, and pathophysiologically grounded alternative to the full GRBAS system, particularly suitable for standardized and reproducible perceptual voice assessment in clinical and research settings. In the Methods section we have added the following explanation to the revised manuscript to improve clarity for international readers (2.3. Examiners' Tools and Standards, 2nd para, page 6, line 221 ff):

"Owing to its validated reliability, focus on the core perceptual dimensions of voice quality (grade, roughness, and breathiness), and well-established pathophysiological basis — with roughness reflecting aperiodic vocal fold vibration and breathiness reflecting incomplete glottal closure — the GRB scale provides an efficient and robust tool for auditory-perceptual assessment."

4. The term "VHIs" is unclear—does this refer to the 30-item questionnaire? This should be explicitly stated. The manuscript notes that VHI-9 is a 9-item instrument, but the distinction between VHI, VHIs, and VHI-9 needs clarification.

We agree to improve clarity by specifying that the overall rating (VHIs) represents a single-item summary of the patient's current vocal impairment and is distinct from the total score of the VHI-9i. We therefore added the following explanatory clause (2.3. Examiners' Tools and Standards, 2nd para, page 6, line 237 ff):

"In a summarizing statement, patients had to assess their overall self-perceived vocal impairment (VHIs) at the present time of questioning on a scale from 0 to 3 (0 = normal, 1 = mild, 2 = moderate, 3 = severe), representing a single-item overall rating separate from the VHI-9i total score."

The rest of the revised section already clearly distinguishes between the original VHI, its short forms, and the VHI-9i, ensuring that the different instruments and scores are properly separated and understandable for the reader.

5. Please include the VHI-9 questionnaire in the appendix with appropriate references.

We thank the reviewer for the suggestion and included the VHI-9i and VHIs in the appendix of the revised manuscript, along with appropriate references (Nawka et al., 2009; Caffier et al., 2021), to facilitate reader access and ensure transparency. The following sentence was added (2.3. Examiners' Tools and Standards, 2nd para, page 6, line 240 f):

"The full VHI-9i and the VHIs are provided in the Appendix, along with relevant references."

Appendix A (page 19, line 703 ff):

**Table A1.** VHI-9i questionnaire items (*German translation*) as used in the present study. The original instrument was developed by Nawka et al. [78], and its validation and classification were described in Caffier et al. [40].

| Item Text | Score | | | | |
|---|---|---|---|---|---|
| My voice makes it difficult for people to hear me. (*Man hört mich wegen meiner Stimme schlecht.*) | 0 | 1 | 2 | 3 | 4 |
| People have difficulty understanding me in a noisy room. (*Anderen fällt es schwer, mich in einem lauten Raum zu verstehen.*) | 0 | 1 | 2 | 3 | 4 |
| The sound of my voice varies throughout the day. (*Der Klang meiner Stimme ändert sich im Laufe des Tages.*) | 0 | 1 | 2 | 3 | 4 |
| My family has difficulty hearing me when I call them throughout the house. (*Meine Familie hört mich kaum, wenn ich zuhause nach ihnen rufe.*) | 0 | 1 | 2 | 3 | 4 |
| My voice difficulties restrict my personal and social life. (*Meine Stimmschwierigkeiten schränken mich in meinem Privatleben ein.*) | 0 | 1 | 2 | 3 | 4 |
| The clarity of my voice is unpredictable. (*Bevor ich spreche, weiß ich nicht, wie klar meine Stimme klingen wird.*) | 0 | 1 | 2 | 3 | 4 |
| My voice is worse in the evening. (*Abends ist meine Stimme schlechter.*) | 0 | 1 | 2 | 3 | 4 |
| I am less outgoing because of my voice problem. (*Ich bin weniger kontaktfreudig wegen meines Stimmproblems.*) | 0 | 1 | 2 | 3 | 4 |
| My voice makes me feel incompetent. (*Wegen meiner Stimme fühle ich mich unfähig.*) | 0 | 1 | 2 | 3 | 4 |

Scoring: 0 = never (*nie*), 1 = almost never (*selten*), 2 = sometimes (*manchmal*), 3 = almost always (*oft*), 4 = always (immer).

**Table A2.** Global VHIs question added to the study questionnaire.

| Question | Score | | | |
|---|---|---|---|---|
| How do you rate your voice today? (*Wie schätzen Sie Ihre Stimme heute ein?*) | 0 | 1 | 2 | 3 |

Scoring: 0 = normal (*normal*), 1 = mildly (*leicht*), 2 = moderately (*mittelgradig*), 3 = severely disturbed (*hochgradig gestört*).

6. A brief explanation of DSI would be valuable for readers who may be less familiar with these measures. I have noted that the authors have explained VEM.

We agree, and integrated the following brief explanation (2.3. Examiners' Tools and Standards, 3rd para, page 5, line 252 ff):

"The DSI is a composite measure calculated as a weighted combination of F0max, SPLmin, MPT, and jitter, providing an overall estimate of dysphonia severity."

7. Section 2.2: Please specify which surgical techniques were used to preserve voice function. Were ELS Type 1 or Type 2 approaches employed? Subepithelial resections correspond to ELS Type 1.

We agree that specifying the surgical approaches is important. The methods section (2.2) describes only the treatment procedure and follow-up protocol. The techniques used to preserve voice function, including subepithelial (ELS Type I) and subligamental (ELS Type II) resections with the actual intraoperative outcomes, i.e., the number and proportion of lesions treated with each approach, are now reported in the Results section (3.2. Intraoperative Assessment, 1st para, page 9, line 337 ff):

"Regardless of the treatment modality, all leukoplakic lesions were successfully excised under microscopic visualization: according to the classification of endoscopic cordectomies of the European Laryngological Society (ELS), type I (subepithelial) resections with complete preservation of the lamina propria were performed in 46 VF lesions (85%), while type II (subligamental) resections including the superficial lamina propria were required in 8 lesions (15%)."

8. Page 4, line 157: Please clarify the rationale for overnight admission, as the described procedure appears to be a minimally invasive excision of a leukoplakic lesion.

We agree to clarify the rationale for overnight admission. Although the procedures were minimally invasive, the brief postoperative stay allowed for careful monitoring of the airway, timely management of any bleeding, and supportive care for pain or other immediate postoperative issues. We added the following information (2.2. Treatment Procedure and Follow-Up Protocol, last para, page 5, line 196 ff):

"Patients were observed on the ward for one or two nights postoperatively as a precautionary measure to monitor airway safety, manage potential postoperative bleeding, and provide immediate support for pain control and early complications."

9. Section 2.3 (Examiners' Tools and Standards): How was voice recording standardised?

– What was the mouth-to-microphone distance and how was background noise controlled?

We agree to address this important information and added the following 2 sentences (2.3. Examiners' Tools and Standards, last para, page 6, line 244 ff):

"Recordings were performed in a sound-treated room (voice laboratory) of our department, with background noise levels below 40 dB(A), and a constant mouth-to-microphone distance of 30 cm maintained using a head-mounted microphone. This recording distance was selected to reduce proximity effects and plosive-related pressure artifacts while controlling for room reflections, ensuring reliable acoustic signal capture, and was supported by automatic SPL calibration within the system."

– Did the phoniatricians assess recorded samples, or did they evaluate patients live in clinic?

We agree that this information was not sufficiently clear from the manuscript and have now specified it as follows (2.3. Examiners' Tools and Standards, 2nd para, page 6, line 217 f):

"Voice samples were recorded live in the clinical setting and subsequently analyzed."

10. The cohort includes 34 unilateral and 10 bilateral cases. What were the voice outcomes specifically for patients undergoing bilateral vocal fold surgery? Apologies if this information is present but was not easily identifiable.

We thank the reviewer for this relevant question. Voice outcomes were analyzed for the entire cohort rather than being stratified by unilateral versus bilateral vocal fold involvement. Therefore, specific outcome data for patients undergoing bilateral vocal fold surgery are not reported separately. However, bilateral cases were included in the overall analysis, and no distinct subgroup effects were identified in the evaluated parameters. We agree that a dedicated subgroup analysis may provide additional insights and will consider this in future studies. We added this limitation in the Methods and Limitations sections as follows:

2.4. Data Processing and Analysis (3rd para, page 7, line 274 ff):

"Analyses were performed according to the respective unit of interest, explicitly clarifying the unit of analysis for each endpoint. Lesion-level analyses were applied when outcomes referred to individual VF lesions, whereas patient-level analyses were used for patient-based variables. In cases of bilateral lesions, no statistical adjustment for within-patient clustering was conducted."

4.5. Limitations and Future Directions (page 17, line 655 ff):

"In cases of bilateral lesions within a single patient, statistical independence cannot be fully assumed, which may introduce bias in lesion-based analyses. As the study was not designed for clustered or hierarchical modeling, this limitation should be considered when interpreting the findings."

11. Line 229: Nearly one quarter of cases (24%, n=13) exhibited preserved phonatory mobility and regular mucosal wave propagation. This suggests that the surgical technique was precise enough to maintain normal mucosal wave function in a significant subset of patients.

The observation that 24% of cases exhibited preserved phonatory mobility and regular mucosal wave propagation is consistent with our baseline VLS findings, where a subset of patients demonstrated normal vibratory function despite leukoplakic lesions. This likely reflects smaller, more superficial lesions without involvement of the lamina propria, which are known to allow stable epithelial regeneration without scarring after resection. Moreover, the overall postoperative improvement, reaching 76% normal vibratory function ($p < 0.001$) supports that the applied phonomicrosurgical techniques were generally tissue-sparing and effective in preserving or restoring mucosal wave function.

12. Section 3.2: Please clarify why some patients underwent laser surgery while others received cold-steel excision.

We agree that clarifying the rationale for the choice of surgical technique improves transparency. The decision between laser-assisted and cold steel excision was not based on specific lesion characteristics, but rather on practical and individualized factors, including the availability of laser systems, surgeon preference and experience, and, in selected cases, patient preference as discussed during the preoperative informed consent process. No systematic bias was observed indicating that lesion localization, size, or morphology influenced the selection of the surgical modality. Both techniques were considered technically equivalent for the intended subepithelial resections, which constituted the majority of cases (85%, ELS type I), and were applied using comparable surgical principles. In addition, specific intraoperative considerations, such as the hemostatic advantages of the blue laser, may have influenced decision-making in individual cases. Overall, the distribution of techniques reflects routine clinical practice rather than protocol-driven allocation, particularly as no significant differences in vocal outcomes were observed between modalities.

We added the following (3.2. Intraoperative Assessment, 1st para, page 8, line 324 ff):

"The choice of surgical modality was not determined by specific lesion characteristics, but rather by pragmatic and individualized factors, including the availability of laser systems (e.g., due to parallel use in other procedures), surgeon preference and experience, and, in selected cases, patient preference discussed during informed consent. Both approaches were considered technically equivalent for the intended subepithelial resections, and no bias was observed indicating that lesion size, localization, or morphological features influenced the selection of the surgical technique."

13. Line 321: All parameters except DSI and F0max showed slight average improvements. Is there an explanation for this pattern?

Yes, the lack of improvement in both DSI and F0max can be explained based on the study results and the preoperative status. Regarding F0max, many patients already exhibited relatively preserved phonatory function at baseline (24% with normal mobility), suggesting that maximum pitch may have been close to physiological limits preoperatively. Additionally, in more extensive lesions, partial involvement or resection of the superficial lamina propria may limit the ability to generate maximal vocal fold tension, thereby constraining increases in highest pitch despite overall functional recovery.

For the DSI, it should be noted that F0max is one of its constituent parameters, so the absence of improvement in F0max directly limits changes in the composite score. Moreover, the DSI integrates several objective measures (jitter, MPT, SPLmin), which in this cohort showed only minor, non-significant changes, thereby reducing the likelihood of a measurable overall increase. Finally, given the predominantly mild to moderate preoperative dysphonia, the potential for detectable postoperative improvement in a composite index such as the DSI was inherently limited.

This explanation is now addressed later in the Discussion, see comment 15.

14. Table 2: Please clarify the distinction between VHI and VHIs. Are these not both measures of self-perceived voice handicap? Should these be labelled VHI-9 and VHI-30 (or VHI-10)?

Yes, in both rows presented in Table 2, the parameters VHI and VHIs reflect measures of subjectively perceived voice handicap. As suggested, we have specified the term VHI more precisely; it refers to the international 9-item Voice Handicap Index (VHI-9i).

The VHIs, however, is not related to the VHI-30 or VHI-10. Rather, it represents an additional single-item question that is completed after the VHI-9i questionnaire. This overall rating reflects a global assessment of the patient's current vocal impairment and is therefore distinct from the cumulative score of the VHI-9i. Please also refer to the Appendix for further clarification.

15. Figure 4: The interquartile range for DSI appears wider postoperatively. This should be addressed in the discussion, as it may have clinical implications.

We agree that the greater postoperative variability in DSI merits discussion, particularly as most other objective parameters showed improvement while DSI and F0max did not (see also your comment 13). Accordingly, we have added the following paragraph to the Discussion section (4.1. Subjective and Objective Voice Outcomes, 2nd para, page 14, line 496 ff):

"Postoperatively, the DSI showed a wider interquartile range than preoperatively, reflecting greater variability in composite outcomes, whereas VEM, MPT, and jitter became more uniform. This likely stems from the DSI's composite nature: its constituent parameters F0max, jitter, MPT, and SPLmin were affected to differing extents across patients. F0max showed no improvement, probably because many patients already had relatively preserved phonatory function at baseline (VLS in 24% with normal phonatory mobility) and, in larger lesions, partial involvement of the superficial lamina propria may limit maximal VF tension. Consequently, no gains in F0max and minor changes in the other parameters constrained overall DSI improvement. Additional factors, including lesion extent, resection depth (ELS type I vs. type II), and individual healing processes, likely contributed to the broader distribution. Clinically, this underscores that while overall vocal function improved, the magnitude of change in composite indices like the DSI varies considerably, highlighting the value of multidimensional assessment."

The inclusion of this additional content has further lengthened the discussion, making it necessary to improve readability through structured subheadings. Therefore, the following headings have been introduced:

- 4.1. Subjective and Objective Voice Outcomes (page 14, line 478 ff)
- 4.2. Role of VEM and Complementary Measures (page 14, line 514 ff)
- 4.3. Surgical Management and Functional Recovery (page 15, line 543 ff)
- 4.4. Histopathology, Malignant Transformation, and Prognosis (page 16, line 574 ff)

- 4.5. Limitations and Future Directions (page 17, line 636 ff)

16. Figure 4: Please expand all abbreviations, including GRB and VHI.

We agree with the reviewer that all abbreviations in Figure 4 (now new Figure 5, following the inclusion of a new Figure 1 showing the flow diagram of patient selection, as requested by Reviewer 2) should be expanded for clarity. Accordingly, we have revised the figure legend as follows (page 13, line 457 ff):

"Middle row: Subjective clinician-rated auditory-perceptual evaluation using the GRB classification (G – overall grade of hoarseness, R – roughness, B – breathiness). Lower row: Subjective self-assessed voice outcomes based on the 9-item voice handicap index (VHI-9i) and overall self-perceived vocal impairment (VHIs) scores."

17. The limitations section is clearly articulated.

Thank you for this positive assessment of the limitations section. We aimed to transparently address key constraints of our study, including the relatively small sample size, retrospective design, and heterogeneity of treatment approaches reflecting real-world clinical practice. In addition, we highlighted the limited follow-up duration and the restriction of postoperative voice assessment to three months, which may not fully capture long-term vocal outcomes or late recurrences. As requested by Reviewer 2, we added explicit acknowledgement of selection bias (44/714 patients), single-center, and lack of adjustment for multiple testing.

4.5. Limitations and Future Directions, page 17, line 637 ff:

"This study has several limitations, particularly the relatively small sample size, which limits statistical power and generalizability. In addition, the final analysis included only 44 of 714 initially screened patients, introducing potential selection bias."

4.5. Limitations and Future Directions, page 17, line 659 ff:

"Eighth, the study was conducted at a single center, limiting generalizability to other settings. Ninth, no adjustment was made for multiple testing, which may increase the risk of type I error."

Thank you for your helpful advice to improve the quality of our manuscript. Furthermore, we revised it according to the recommendations of the second reviewer. We hope that our paper can now be accepted.

30.3.2026 Philipp P. Caffier, for all authors