## Comments to Reviewer 2:

The topic is clinically relevant and the dataset is carefully gathered with thoughtful multidimensional voice assessment. However, there are some issues that need to be addressed.

**Major comments**

1. Unit of analysis & statistical independence: The manuscript blends per-patient and per-lesion counts without making the analytic unit explicit, and some labels appear incorrect. Example: Table 1 is labeled Resection type per patient but totals 30 + 18 + 6 = 54, suggesting per-lesion, not per-patient. Likewise, postoperative phonatory mobility sums to 54, again suggesting lesion-level rather than patient-level tallies. Please correct the labels and ensure consistent counting throughout. If some analyses are lesion-level while others are patient-level, the current statistical approach may violate independence (bilateral lesions within a single patient are not independent). Re-specify the unit of analysis for each endpoint and re-analyze accordingly.

Thank you for this important and insightful comment. We fully agree that a clear specification of the unit of analysis is essential. The table labeling itself was correct; however, there was an inconsistency in Table 1 caused by a last-minute revision prior to submission. Specifically, the row originally labeled “resection type per patient” had already been changed to “resection type per lesion”, but the corresponding heading was not updated accordingly. We have now corrected this by removing “per patient,” such that the table consistently refers to lesion-based counts (n = 54), unless otherwise explicitly stated. The reported numbers and percentages are correct for resection type per lesion; only the label was incorrect due to this late-stage modification before submission.

Table 1 was corrected as follows (3.1. Subjective Demographics and Baseline Assessment, page 8, line 315 ff):

**Before correction:**

| | Number | % |
|---|---|---|
| Resection type per patient | | |
| cold steel dissection | 30 | 55.6% |
| $CO_2$ laser assisted | 18 | 33.3% |
| blue laser assisted | 6 | 11.1% |

**After correction:**

| | Number | % |
|---|---|---|
| Resection type | | |
| cold steel dissection | 30 | 55.6% |
| $CO_2$ laser assisted | 18 | 33.3% |
| blue laser assisted | 6 | 11.1% |

Regarding the statistical analysis, no reanalysis was performed, as analyses were conducted appropriately according to the respective unit of interest: lesion-level analyses were used when referring to lesions, and patient-level analyses when referring to patient-based

outcomes. We acknowledge the reviewer's important point regarding the potential violation of statistical independence in cases of bilateral lesions within the same patient. Given the retrospective nature of the study and the structure of the dataset, reanalysis using more advanced hierarchical or clustered statistical models is not feasible. Therefore, we have addressed this issue by explicitly clarifying the unit of analysis for each endpoint in the Methods section and by adding a corresponding statement to the Discussion section, highlighting the potential lack of independence in lesion-based analyses.

We added in the Methods section (2.4. Data Processing and Analysis, last 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."

The Discussion section was updated to include (4.5. Limitations and Future Directions, page 17, line 654 ff):

"Seventh, some analyses were conducted at lesion level, while others were performed at patient level. 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."

These revisions improve the clarity, transparency, and methodological rigor of the manuscript while also highlighting the existing limitations.

2. You start with 714 adults and end with 44 analyzed; reasons for exclusion are listed narratively but the numbers per exclusion step are not shown. This raises the risk of selection bias and makes reproducibility difficult. Add a flow diagram (e.g., STROBE-style) detailing: screened (714) → excluded with reasons and counts (resolution, minimal lesion, suspected invasive cancer, non-leukoplakia mimickers, declined surgery, incomplete data, lost to follow-up, etc.) → included (n=44).

We thank the reviewer for this helpful suggestion. A flow diagram has now been added (new Figure 1, page 4, line 143 ff, see below), detailing the stepwise selection of patients from the initial cohort (n = 714) to the final study population (n = 44), including reasons for exclusion with corresponding counts. We also added the following sentence to the Methods section (2.1. Study Design and Patients, page 3, line 116 f):

"A flow diagram illustrating patient selection and sequential application of exclusion criteria is shown in Figure 1."

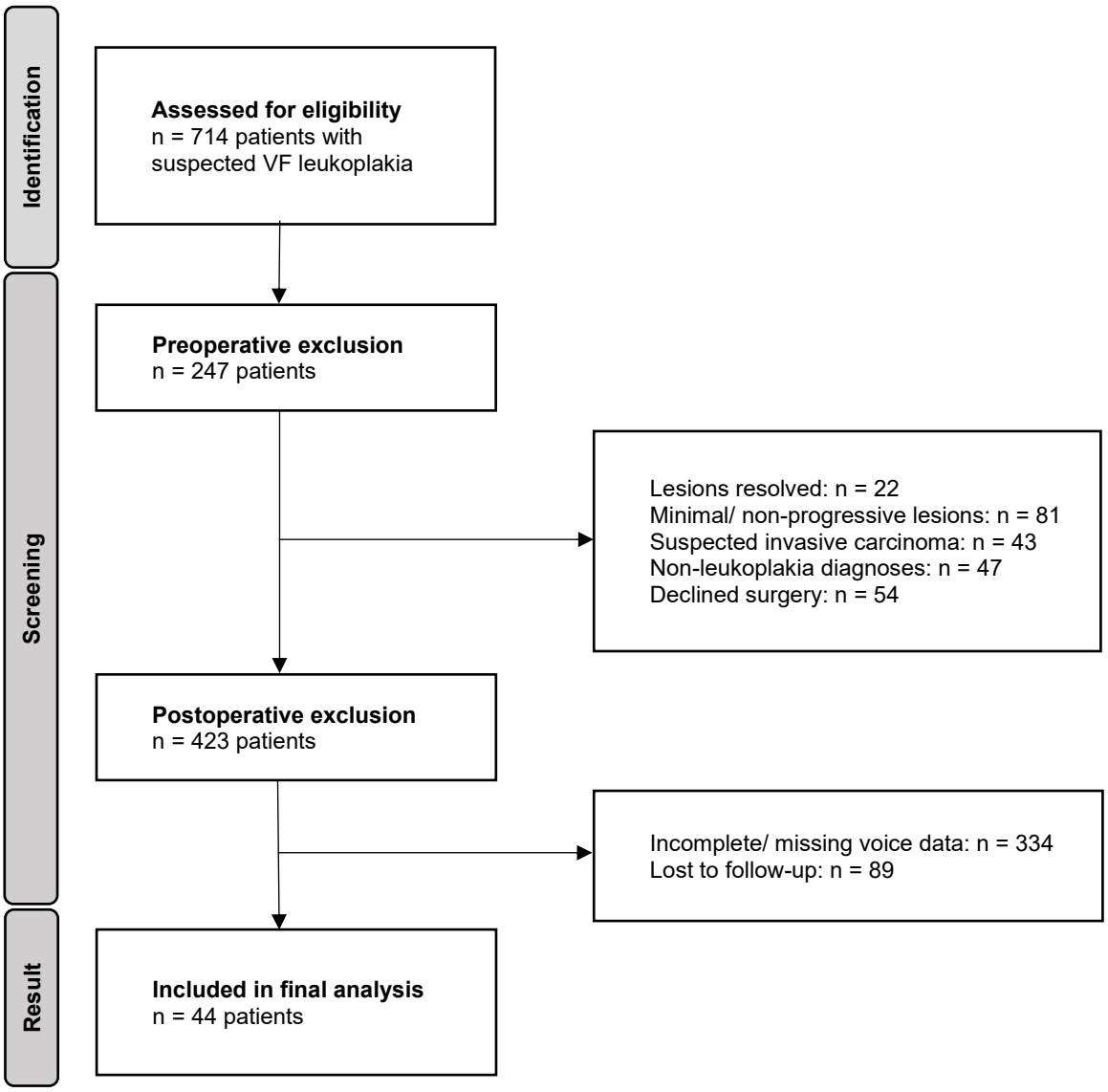



**Figure 1.** Flow diagram of patient selection. Study participants were sequentially excluded based on predefined clinical and methodological criteria. Exclusion categories were applied hierarchically, so that each patient was counted only once, corresponding to the first applicable criterion. Only individuals with complete pre- and postoperative multidimensional voice assessments and histopathological confirmation were included in the final analysis.

3. In the Abstract you state that phonosurgical method, histopathology, and age did not significantly affect voice outcomes, while younger age was associated with greater VEM gains (the latter implies an age effect).

We thank the reviewer for this insightful comment. We agree that the current wording in the Abstract may suggest a contradiction. Importantly, two different analytical perspectives are being reported: (i) overall postoperative voice outcomes and (ii) associations with the magnitude of changes (Δ values). In our analysis, age did not significantly influence absolute

postoperative voice outcomes. However, when analyzing pre–post changes, a trend was observed indicating that younger patients exhibited greater improvements in VEM. This reflects an association with the extent of functional gain rather than with the final outcome level. To avoid ambiguity, we have revised the Abstract to clearly distinguish between these aspects. We replaced “were associated with” with “showed trends toward.” This revision more accurately reflects the data: “did not significantly affect” corresponds to the primary analysis, while “showed trends” appropriately qualifies the age-related VEM finding ($p = 0.095$). The change maintains parallel structure and improves clarity, ensuring that the Abstract is both consistent and statistically accurate without increasing its length. The sentence has been revised as follows (Abstract, page 1, line 37 ff):

| **Before correction:** | **After correction:** |
| --- | --- |
| Phonosurgical method, histopathology, and age did not significantly affect voice outcomes, while higher dysplasia grades and younger age were associated with greater VEM gains. | Phonosurgical method, histopathology, and age did not significantly affect voice outcomes; however, higher dysplasia grades and younger age showed trends toward greater VEM gains. |

4. You report 6/44 (14%) recurrences over a mean of 16 months (range 3–46), but means are not ideal for time-to-event data with variable follow-up.

We thank the reviewer for this important methodological comment. We agree that, in classical time-to-event analyses with standardized follow-up and complete event capture, median values or Kaplan–Meier estimates are generally preferred over means. However, in the present study, follow-up was not protocol-driven beyond the scheduled 2-week and 3-month postoperative visits, after which further consultations were patient-initiated. While all patients received identical follow-up recommendations and were explicitly instructed to return in case of recurrent symptoms, long-term follow-up intervals varied, and complete event ascertainment cannot be guaranteed. This is partly attributable to the real-world nature of our cohort, which includes patients from a wide geographic area. Given this non-standardized and likely incomplete follow-up, the assumptions underlying formal time-to-event analyses are not fully met. In this context, both mean and median values would be subject to similar limitations and should be interpreted descriptively rather than as precise estimates of recurrence dynamics. To improve clarity, we have revised the manuscript to explicitly state the descriptive nature of these data and the limitations of follow-up with regard to recurrence assessment. We added in the Results and Discussion sections:

3.3. Histopathological Results and Surgical Outcomes (page 9, line 357 ff):

“Within the mean postoperative observation period of 16 months (range, 3–46), 6 patients (14%) experienced lesion reappearances; follow-up duration varied and should be interpreted descriptively.”

4.5. Limitations and Future Directions (page 17, line 643 ff):

"Fourth, follow-up duration varied and may have been too short to capture late recurrences or long-term vocal changes. In addition, follow-up beyond the scheduled postoperative visits was not protocol-driven but patient-initiated, which may have led to incomplete detection of recurrence events. Consequently, time-to-event data should be interpreted descriptively, and summary measures such as mean follow-up or recurrence-free intervals are subject to inherent limitations."

5. All postoperative voice measures were obtained at three months only; please justify this time point and, if possible, provide information on longer-term voice outcomes (even descriptively) for those with extended follow-up. Discuss clinical significance beyond statistical significance.

Postoperative voice assessments were performed at three months to capture complete wound healing and optimal vocal recovery after phonomicrosurgery, which aligns with standard practice in functional laryngeal evaluation. Patients with longer-term follow-up did so voluntarily, either for routine reassessment or due to perceived changes in their voice to rule out recurrence. Those who underwent extended multidimensional voice diagnostics (typically annually) demonstrated stable voice outcomes, suggesting that, in the absence of recurrence, vocal function remains largely preserved over time.

While statistical significance captures measurable improvements, the clinical relevance is underscored by consistent reductions in VHI-9i scores, improved GRB ratings, and enhanced VEM values, reflecting perceptible and meaningful benefits to patients' everyday voice use. These changes are thus both statistically and functionally significant.

We integrated these aspects into the Methods, Results and Discussion sections as follows:

2.2. Treatment Procedure and Follow-Up Protocol (last para, page 5, line 202 ff):

"This 3-month interval allowed sufficient time for complete wound healing and functional recovery. Patients with longer-term follow-up underwent additional assessments either as part of routine reevaluation or in response to perceived changes in voice. In some cases, extended follow-up evaluations were performed annually to assess long-term outcomes and to examine whether postoperative voice improvements were maintained over time in the absence of recurrence."

3.4. Objective and Subjective Vocal Outcomes (last para, page 13, line 450 f):

"In patients with extended follow-up, voice outcomes remained stable over time."

4.1. Subjective and Objective Voice Outcomes (3rd para, page 14, line 509 ff):

"Although our main postoperative assessment was at three months, extended follow-up in a subset of patients indicates that voice improvements are maintained over time. Beyond statistical significance, these improvements are clinically meaningful, reflected in lower VHI-9i scores, improved GRB ratings, and enhanced VEM, reflecting perceptible benefits in everyday voice use."

4.5. Limitations and Future Directions (page 17, line 650 ff):

"Sixth, postoperative voice assessment was restricted to three months, preventing formal evaluation of long-term adaptation. Although some patients underwent extended follow-up assessments, these were not systematically collected, limiting the ability to draw robust conclusions on long-term voice trajectories despite indications of sustained functional stability."

**Minor comments**

1. Table 1 labeling & structure: Split per-patient and per-lesion summaries into separate tables or subpanels; ensure headers match counts.

Thank you for this helpful suggestion. We agree that a clear distinction between patient- and lesion-level data is important. However, we chose not to split Table 1 into separate tables, as the current structure – presenting preoperative (left) and postoperative (right) data side by side – provides a direct and clinically meaningful comparison for the reader. Separating these into different tables would reduce this comparability and overall readability.

To address the concern and improve clarity, we have revised Table 1 as follows:

- On the left side, only two variables refer to patient-level data (Gender and Age). To make this explicit, we added "patients" to the age category so that both variables are presented consistently and clearly as "male patients" and "female patients":

**Before correction:**

| | | |
|---|---|---|
| Gender | | |
| male patients | 29 | 65.9% |
| female patients | 15 | 34.1% |
| Age [years]; M (R) | | |
| male | 63 (30-80) | |
| female | 58 (37-81) | |

**After correction:**

| | | |
|---|---|---|
| Gender | | |
| male patients | 29 | 65.9% |
| female patients | 15 | 34.1% |
| Age [years]; M (R) | | |
| male patients | 63 (30-80) | |
| female patients | 58 (37-81) | |

- On the right side, patient-level information is limited to "Patient treatment response" and "Time-related aspects". To avoid ambiguity, the latter was renamed to "Patient time course", aligning it with the structure of "Patient treatment response" and clearly indicating that these data refer to patients rather than lesions. As both aspects are presented consecutively, the reader can follow the information without unnecessary switching.

**Before correction:**

| | | | |
|---|---|---|---|
| Patient treatment response | | | |
| complete primary resection | 42 | 95.5% | 95.5% |
| second look surgery (d/t path) | 2 | 4.5% | 4.5% |
| recurrence | 6 | 13.6% | 13.6% |
| Time-related aspects; M (R) | | | |
| duration until diagnosis [days] | 33 (1-172) | | |
| follow-up period [months] | 16 (3-46) | | |
| time until recurrence [months] | 19 (3-40) | | |

**After correction:**

| | | |
|---|---|---|
| Patient treatment response | | |
| complete primary resection | 42 | 95.5% |
| second look surgery (d/t path) | 2 | 4.5% |
| recurrence | 6 | 13.6% |
| Patient time course; M (R) | | |
| duration until diagnosis [days] | 33 (1-172) | |
| follow-up period [months] | 16 (3-46) | |
| time until recurrence [months] | 19 (3-40) | |

With these adjustments, all variables in the table are now clearly attributable to either the patient or lesion level, and potential sources of confusion have been resolved. We therefore believe that the current table structure can be maintained without compromising clarity or interpretability.

2. Abstract: Replace post vs. pre with post vs preoperative for clarity.

We agree and have expanded the abbreviations as suggested (Abstract, page 1, line33 f):

“Postoperatively, subjective measures showed significant improvements (post- vs. preoperative), with decreased VHI-9i scores (10 vs. 14) and GRB ratings ($p < 0.05$).” th

3. Limitations: You provide a good list; please add explicit acknowledgement of selection bias (44/714), single-center, and lack of adjustment for multiple testing.

We agree and added (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.”

Furthermore (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 first reviewer. We hope that our paper can now be accepted.

30.3.2026 Philipp P. Caffier, for all authors