arXiv:2602.02725v1 [cs.LG] 2 Feb 2026

使用非侵入式颈部声学传感进行自动化吞咽障碍筛查

Jade Chng[1, 2, *], *Rong Xing*[1, *], *Yunfei Luo*[3, *], *Kristen Linnemeyer-Risser*[4], *Tauhidur Rahman*[1,3], *Andrew Yousef*[4, †], *Philip A Weissbrod*[4, †]

[1] 雅各布斯工程学院，加利福尼亚大学圣迭戈分校[2] 生物医学工程系，杜克大学[3] Halıcıoğlu数据科学研究所，加利福尼亚大学圣迭戈分校˙耳鼻咽喉头颈外科，加利福尼亚大学圣迭戈分校[4]

摘要

咽部健康在呼吸、吞咽和发声等人类基本功能中发挥着至关重要的作用。及早检测吞咽异常（也称为吞咽困难）对于及时干预至关重要。然而，当前的诊断方法通常依赖放射影像学检查或侵入性操作。在本研究中，我们提出了一种自动化框架，结合应用机器学习与便携式无创声学传感，用于检测吞咽困难。通过在吞咽任务期间采集颈部的细微声学信号，我们旨在识别与异常生理状况相关的模式。在5次独立训练-测试划分下，我们的方法在测试时异常检测方面取得了良好表现，AUC-ROC达到0.904。本研究证明了采用无创声学传感作为咽部健康监测的实用且可扩展工具的可行性。

索引词— 应用机器学习、信号处理、咽部健康、数字健康。

1. 引言

吞咽困难，即吞咽障碍，仍然是一个重大的公共卫生问题，影响着50岁以上成年人中的10%至20%，以及多达三分之二的帕金森病、卒中或头颈癌等疾病患者 [1, 2]。据估计，每年约有43亿至71亿美元用于与吞咽困难相关的住院费用 [3]。目前，吞咽评估通过基于视频的检查方式进行，例如柔性内镜吞咽评估（FEES）、电视透视吞咽检查（VFSS）或高分辨率测压。这些方法虽然能够提供有价值的信息，但要么具有侵入性，要么需要接受辐射，要么必须由经过培训的专业人员操作，或者在成本效益方面存在局限 [4, 5]。鉴于当前金标准吞咽评估存在局限，具有误吸风险的住院患者通常会接受筛查以进行误吸筛查，通常采用临床吞咽评估。然而，这些评估在未使用仪器的情况下进行，诊断敏感度和特异度有限，并且常常使医务人员在临床决策方面感到不确定。这凸显了对一种客观吞咽评估工具的迫切需求，该工具能够高效、准确地识别吞咽障碍风险。

为弥补这一不足，我们在进行纤维内镜吞咽评估（FEES）期间收集了听力测量数据，以研究吞咽功能障碍与声学特征之间的关联。随后，我们利用这些数据开发了一种用于预测吞咽功能障碍的机器学习模型。我们提出的系统取得了良好的分类性能，这表明表面数字听诊能够捕捉到具有诊断价值的声学特征标记，反映潜在的口咽功能障碍。

2. 相关工作

为应对吞咽困难评估带来的这些挑战，越来越多的研究也开始探索运用机器学习技术，结合加速度测量、声学信号和肌电图（EMG）来分析吞咽机制并检测相关异常 [6,7, 8, 9]。尽管初步研究表明，机器学习有望用于检测吞咽功能障碍，但这些研究仍存在显著局限。首先，最值得注意的是，尚无研究在柔性内镜吞咽评估（FEES）等金标准评估期间实时评估吞咽声音，因此难以描述功能障碍的严重程度和性质。其次，许多研究仅使用单一食团稠度（例如一口水）验证模型 [6, 7, 8, 9], , 尽管吞咽困难和误吸在不同稠度下可能表现各异，这限制了研究的外部效度。最后，大多数既往研究局限于样本量较小的概念验证设计 [10]。为克服这些局限，我们开发了一种机器学习模型，该模型使用全面吞咽评估期间采集的声学数据进行训练，数据涵盖了600多次吞咽事件。

*本文作者对本研究的贡献相同。
†共同资深作者

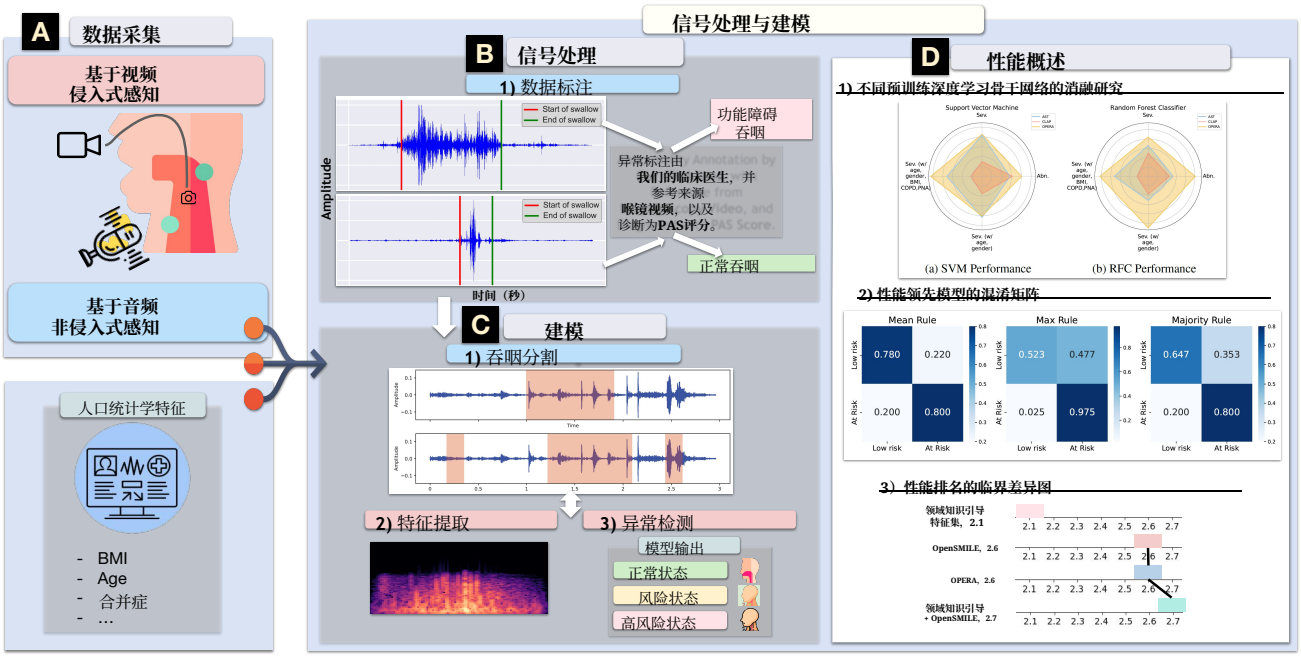


图1：所提出的自动化系统概览。(A) 本研究期间从参与者处收集的数据。(B) 数据标注流程示例。(C) 用于探索声学信号与吞咽异常之间关系的建模过程。(D) 实证结果展示。

3. 方法

3.1. 初步实验

我们首先进行了初步实验，考察已有基线音频嵌入模型AST、CLAP和OPERA的性能 [11, 12, 13] （图1第D.1节）。实验使用现有数据集，并在有无人口统计特征的情况下进行，采用了5组随机训练-测试划分。“Sev.”和“Abn.”分别代表严重程度（3分类）和异常（2分类）。结果显示，OPERA模型性能最佳，因此我们将其纳入主要实验，作为基线进行比较。

尽管年龄和性别对模型性能的提升并不显著，但鉴于先前研究表明年龄和性别可能改变吞咽的声学特征 [14]，我们决定将其纳入。此外，RFC和支持向量机的性能相当，但为保持一致性，我们选择在主要评估中使用RFC。

3.2. 数据标注

本研究经加利福尼亚大学圣迭戈分校批准，我们从气道、嗓音与吞咽中心共招募了49名自我报告有吞咽困难症状的参与者。这些参与者接受了使用柔性内镜吞咽评估（FEES）进行的吞咽评估。标准的柔性内镜吞咽评估包括8至10次不同稠度和食团大小的经口摄入试验。在此过程中，将柔性喉镜经鼻置入，并观察每一次吞咽试验。言语语言病理学家和接受过专科培训的喉科医师使用渗透-误吸量表（PAS）对每次柔性内镜吞咽评估进行评分，以评估吞咽功能障碍的严重程度。PAS是一种

表1：参与者背景分布

因素	样本量 (n)	百分比 (%)
性别		
女性	25	51.02
Male	24	48.98
PAS		
正常		
PAS[1-2]	22	44.90
异常		
PAS[3-5]	20	40.82
PAS[6-8]	7	14.29
总计	27	55.10
BMI		
< 18.5(体重过轻)	2	4.08
≤ 24.9(正常)	17	34.69
> 24.9(超重)	30	61.22
Age		
30–60	13	26.53
62–69	16	32.65
70–96	20	0.41
慢性阻塞性肺疾病（COPD）		
No	41	83.67
Yes	8	16.33
误吸所致肺炎（PNA）病史		
No	38	77.55
Yes	11	22.45

分类评估评分，范围为1–8，用于评估功能障碍程度以及经口摄入物侵入气道的证据，分数越高表示吞咽功能障碍程度越高。根据该量表，PAS为1-2属于相对正常的吞咽，PAS为3-5表明存在喉前庭穿透的证据，PAS为6及以上则表明存在误吸的证据。在柔性内镜吞咽评估期间，每位参与者还在甲状软骨外侧放置数字听诊器（3MTM Littmann Core DigitalStethoscope），以便在评估过程中实时采集听力测量数据。

本研究共收集了 392 段音频录音。我们剔除了损坏的音频文件，这些文件中包含说话声或非吞咽声音，且无法通过参数滤除。随后，对每段录音中的吞咽声音进行清理，以便进行特征提取；清理后的平均时长为 0.64 秒。处理后，共获得 617 个独立的

吞咽事件（24 名参与者各贡献了 10–15 个吞咽事件，10 名参与者各贡献了 15–20 个事件，8 名参与者各贡献了 $\leq$10 个事件，3 名参与者各贡献了 $\geq$20 个事件）。我们的目标是将基于领域知识的特征与从 OpenSMILE 提取的特征以及 OPERA 的嵌入向量进行比较，以区分存在吞咽功能障碍的患者和不存在吞咽功能障碍的患者，并可能反映患者病情的严重程度。患者背景数据的分布见表 1。

我们使用 Python Librosa [15] 处理并“清理”本研究收集的音频录音，并使用若干参数来准确识别每个音频文件中所有吞咽的起始和结束位置（即单个片段）。这些参数包含 4 个要素：(i) 以 dB 为单位的振幅阈值，用于确定静音时段；(ii) 以秒为单位的间隔时间，在此范围内的片段将合并为一个片段；(iii) 最小振幅，振幅低于此值的片段将被舍弃；以及 (iv) 最大振幅，振幅高于此值的片段将被舍弃。通过与吞咽的视觉数据交叉比对，我们针对每个音频文件分别调整参数，从而确保所有吞咽都被正确分段（图1 B.1 节）。

3.3. 特征提取与建模

除了提取 OPERA嵌入外，我们还分析了一系列在启发式层面上与吞咽功能障碍相关的领域知识引导特征 [16]。

3.3.1. 频率

为了计算每个音频文件中最大的五个频率，我们使用了快速傅里叶变换。该算法高效地实现了离散傅里叶变换（DFT）方程，将信号从时域 $x[n]$ 转换到频域 $X[k]$。DFT 方程：

$$X[k] = \sum_{n=0}^{N-1} x[n]\, e^{-j2\pi kn/N} \tag{1}$$

其中， $x[n]$ 是时间索引 n处的输入信号， j 是虚数单位， k 是频率箱的索引， N 是样本总数， $X[k]$ 是变换后的频域信号。

随后，我们使用 Librosa 的短时傅里叶变换（STFT）函数，将此公式反复应用于重叠窗口，从而获取音频文件的完整频率成分。根据这些随时间变化的频率成分，我们计算了每个音频文件随时间变化的平均频率和中位数频率。最后，我们提取了最大的五个频率值： $freq_i$ 其中 $i = 1, \ldots\ldots, 5$。

3.3.2. 振幅

为获取与振幅相关的数值，我们使用了 Librosa 的 load 函数，将 .WAV 音频文件转换为由浮点值构成的一维时间序列数组。这些数值分别代表文件中所有采样点的振幅。因此，我们利用该数组计算了随时间变化的峰值（最大值）振幅和平均振幅。

3.3.3. 曲线下面积T

o calculate the area under the curve, we utilized the NumPy库提供的复合梯形法则对绝对波形进行积分。复合梯形法则的显式公式为：

$$\int_a^b f(x)\,dx \approx \frac{1}{2}\sum_{j=1}^{n}(x_j - x_{j-1})\,[f(x_{j-1}) + f(x_j)] \tag{2}$$

其中 a, b 为积分区间， n 为分区数。

3.3.4. 来自预训练音频模型的表示。

最后，为了与文献保持一致，我们还收集了 OpenSMILE 的 [17] 工具包提供的所有特征。（对于这三种特征提取方法，均纳入了参与者的年龄和性别。）

3.4. 分段与聚合分析

在临床环境中，音频录音通常会在单次试验中包含多次吞咽。因此，模型需要从连续输入中检测并分类每次单独的吞咽。为模拟这一情况，我们还通过固定参数 和滑动窗口自动分段。 两种方法评估模型性能。这些方法模拟了在真实部署环境中的自动化预处理。生成逐次吞咽预测后，我们采用三种策略对该患者的所有吞咽结果进行聚合，从而得出患者级别的分类：平均风险（预测均值）、最大风险（所有吞咽中的最高风险）和众数风险（出现频率最高的预测）。

3.4.1. 固定参数分割

通过测试不同阈值组合并最大化与真实掩码的交并比（IoU），确定固定参数。最佳重叠度（IoU = 0.4775）采用顶部 db = 20、间隔时间 = 0.6、允许的最大振幅= 2和允许的最小振幅= 0时获得，灵敏度为 65.8%，特异度为 87.6%。分段示例见图1 C.1 节，其中真实标注位于预测分割上方。

3.4.2. 滑动窗口分割

滑动窗口将音频分割为 1 秒的窗口，重叠率为 50%，确保覆盖整个音频时间轴。这种策略可最大限度地降低漏掉吞咽动作的风险，而且算法简单，适合实时使用。

3.5. 评估方案

所有实验均采用患者级训练测试划分，以更好地反映临床使用场景。这样可以确保模型不会对训练期间已见过的患者过拟合

表 2：主要结果（患者级划分）。“严重程度”和“异常”分别指严重程度和异常。

任务	方法	AUC ROC	AUC PRC	平衡准确率
Sev.	OPERA	0.557 ± 0.159	0.434 ± 0.130	0.542 ± 0.047
Sev.	OpenSMILE (OpSL)	0.583 ± 0.120	0.503 ± 0.145	0.606 ± 0.079
Sev.	领域知识引导	**0.611 ± 0.055**	**0.519 ± 0.061**	**0.659 ± 0.028**
Sev.	领域知识引导 w/ OpSL	0.561 ± 0.135	0.493 ± 0.120	0.610 ± 0.080
Abn.	OPERA	0.651 ± 0.176	0.718 ± 0.140	0.579 ± 0.080
Abn.	OpenSMILE	0.778 ± 0.144	0.850 ± 0.094	0.665 ± 0.152
异常	领域知识引导	**0.904 ± 0.015**	**0.913 ± 0.075**	**0.755 ± 0.061**
异常	领域知识引导（含 OpSL）	0.804 ± 0.183	0.862 ± 0.081	0.710 ± 0.159

表3：采用音频分段的评估，按患者汇总（ROC曲线下面积得分）

方法	Mean	Max	Mode
滑动窗口	0.893 ± 0.103	0.856 ± 0.106	0.884 ± 0.104
固定参数	0.868 ± 0.142	**0.942** ± 0.051	0.842 ± 0.141
人工分割的吞咽	**0.967** ± 0.054	0.918 ± 0.079	**0.971** ± 0.041

。这模拟了我们的目标临床用例，即预期模型能够在完全未见过的患者身上进行验证和推断。我们进行了五次独立的患者级划分，并将患者ID用作数据划分的主要单位。每次划分均经过分层处理，以保持类别分布和吞咽分布。我们报告ROC曲线下面积得分，作为比较模型性能的主要指标。

4. 结果

4.1. 特征性能比较

使用我们的领域知情特征训练的二分类（异常检测）模型表现最佳，其**AUC- ROC 为 0.904**，优于预训练音频嵌入（表 2）。在二分类设置下，将 OpenSMILE 与我们的领域知情特征结合后，模型性能下降，AUC-ROC 得分为 0.804。这是合理的，因为 OpenSMILE 的声学特征可能会捕捉到无关或含噪特征。总体而言，三分类（严重程度分类）设置下的性能明显低于二分类设置。这主要是由于每个类别的训练样本较少，阻碍了模型学习各类别判别模式的能力。进一步使用规模更大、类别分布更均衡的数据集进行探索，或有助于释放多类吞咽分类的潜力。

4.2. 分段与患者聚合结果

人工分割的吞咽条件作为基准，代表所有吞咽均被

准确分割（表3）。固定参数分割方法取得了令人鼓舞的结果，在最大风险聚合策略下观察到最高的 AUC-ROC 分数 0.942。然而，我们注意到，在平均风险和众数风险聚合下，滑动窗口分割优于固定参数方法。这可能表明，尽管固定参数与最大风险组合的结果表现出色，但它可能只是特定数据集带来的偶然结果，而非始终更优的配置。

混淆矩阵采用人工分割的吞咽数据，表明该模型能够有效识别高风险患者，且假阴性极少（图1 D.2 节）。这在临床上具有重要意义，因为将高风险患者误分类，可能导致未能发现存在不安全吞咽的患者，并允许其经口饮食，从而使其面临吸入性肺炎的风险。

4.3. 特征重要性与统计分析

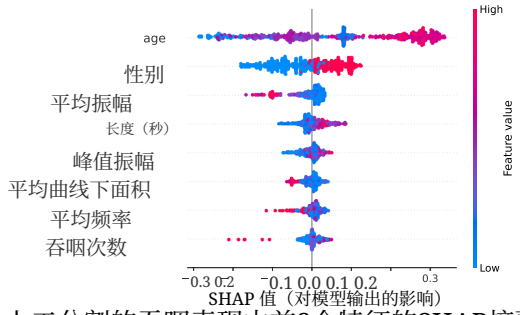


图2：人工分割的吞咽表现中前8个特征的SHAP摘要图

SHAP分析（图2）显示，年龄和性别是吞咽困难的重要预测因素；年龄较大和男性与较高风险相关，这与既往研究一致。[18, 19] 基于信号的特征——平均振幅、频率、曲线下面积和吞咽次数——也显著影响分类，较弱的吞咽与吞咽困难相关。这些发现表明，结合人口统计学特征和声学特征能够最大限度地提高预测准确性。

最后，我们将使用领域知识引导特征的方法与先前构建的声学数据集（OPERA和OpenSMILE）的准确性进行了比较，发现我们基于领域知识引导的方法优于所有其他方法，这一点由临界差异检验所证实（图1 D.3节）。

5. 结论

本研究提出了一种利用无创声学传感检测吞咽困难的机器学习框架。通过采集颈部区域声学信号并运用时频分析，我们的系统展现了高精度识别吞咽异常生理模式的潜力。所提出的方法为传统诊断方法提供了一种低成本、可扩展的替代方案，并代表着迈向便捷实时咽部健康监测的一步。

局限性与未来工作。 尽管前景可期，数据集规模仍然适中，这限制了其在不同人群和状况中的泛化能力。扩大数据集、改进自动吞咽分割算法，并在多种场景（如居家环境）中进行验证，将是下一步的必要工作。

6. 致谢

我们感谢加利福尼亚大学圣迭戈分校医学院的医学生 Sacha Moufarrej、Bao Luu 和 David Zeng 协助拼接吞咽音频文件。本研究未获得外部资助。作者声明不存在相关的财务或非财务利益冲突。所有数据均作为一项获批研究的一部分从患者处收集，并在分析前进行了完整匿名化。

7. 参考文献

[1] Z. W. Seo、J. H. Min、S. Huh、Y.-I. Shin、H.-Y. Ko 和 S.-H. Ko，“采用吞咽视频透视检查评估吸入性肺炎患者吞咽困难的患病率和严重程度”，*Lung*，第 199 卷，第 55–61 页，2021 年。

[2]M. C. Rivelsrud、L. Hartelius、L. Bergstr¨om、M. Løvstad 和 R. Speyer，“不同医疗环境中成年人口咽性吞咽困难的患病率：系统综述与荟萃分析”，吞咽困难，第 38 卷，第 1 期，76–121 页，2023 年。[3] D.A. Patel、S. Krishnaswami、E. Steger、E. Conover、M. F Vaezi、M. R. Ciucci 和 D. O. Francis，“美国住院患者吞咽困难造成的经济负担与生存负担”，食管疾病，第 31 卷，第 1 期，页码 dox131，2018 年。[4] G. Sanjeevi、U. Gopalakrishnan、R. Pathinarupothi 和 K. Iyer，“人工智能在吞咽视频透视检查分析中的应用：综合综述”，吞咽困难，第 1–14 页，2025 年。[5] D. Wong、J. Wang、S. Cheung、D. Lai、A. Chiu、D. Pu、J. Cheung 和 T. Kwok，“吞咽困难筛查的当前技术进展：系统性范围综述”，医学互联网研究杂志，第 27 卷，页码 e65551，2025 年。[6] M. Nikjoo、C. Steele、E. Sejdi´c 和 T. Chau，“使用基于信誉的分类器自动区分安全与不安全吞咽”，在线生物医学工程，第 10 卷，1–18 页，2011 年。[7] C. Steele、R. Mukherjee、J. Kortelainen、H. P¨ol¨onen、M. Jedwab、S. Brady、K. Theimer、S. Langmore、L. Riquelme、N. Swigert 等，“为口咽性吞咽困难风险患者开发无创吞咽筛查设备：前瞻性探索性研究结果”，吞咽困难，第 34 卷，698–707 页，2019 年。[8] C. Steele、E. Sejdi´c 和 T. Chau，“使用双轴吞咽加速度测量无创检测稀液体误吸”，吞咽困难，第 28 卷，105–112 页，2013 年。

[9] H. Mohammadi、A. Samadani、C. Steele 和 T. Chau，“自动区分咳嗽与非咳嗽加速度测量信号伪影”，生物医学信号处理与控制，第 52 卷，394–402 页，2019 年。[10] D. Jayatilake、T. Ueno、Y. Teramoto、K. Nakai、S. Hi-daka、K.and Ayuzawa、K. Eguchi 和 A. Matsumura，“基于智能手机的实时吞咽能力评估：通过吞咽声音进行评估”，*IEEE*健康与医学转化工程期刊，第 3 卷，2015 年 11 月。[11] Y. Gong、Y. Chung 和 J. Glass，“AST：音频频谱图变换器”，2021 年。[12] B. Elizalde、S. Deshmukh、M. Al Ismail 和 H. Wang，“CLAP：通过自然语言监督学习音频概念”，2022 年。[13] Y. Zhang、T. Xia、J. Han、Y. Wu、G. Rizos、Y. Liu、M. Mosuily、J. Chauhan 和 C. Mascolo，“迈向开放式呼吸声学基础模型：预训练与基准测试”，2024 年。[14]S. Youmans 和 J. Stierwalt，“不同年龄、性别、食团黏度和食团体积下的正常吞咽声学特征”，吞咽困难，第 26 卷，第 4 期，374–384 页，2011 年 1 月。[15]B. McFee、M. McVicar、D. Faronbi、I. Roman、M. Gover、S. Balke、S. Seyfarth、A. Malek、C. Raffel、V. Lostanlen、B. van Niekerk、D. Lee、F. Cwitkowitz、F. Zalkow、O. Nieto、D. Ellis、J. Mason、K. Lee、B. Steers 和 D. S¨udholt，“librosa/librosa：v0.11.0”，2025 年，使用 Python 进行音频与音乐信号处理。[16] A. Yousef、S. Moufarrej、D. Zeng、B. Luu、R. Xing、J. Crook、Y. Luo、K. Linnemeyer-Risser、T. Rahman 和 P. Weissbrod，“探听吞咽困难：使用数字听诊器评估吞咽的听力测定差异”，喉镜，2025 年。[17] Oana Pascu、Diana Oneaţˇa、Horia Cucu 和 Nikolaus M¨uller，“简单、可解释、高效：OpenSMILE 在语音深度伪造检测中的应用”，载于*IEEE*国际声学、语音与信号处理会议论文集（*ICASSP*），2025 年，第 1–5 页。[18] C. Bolinger，“肺炎：年龄或性别是否与老年患者吞咽困难的存在有关？”，肺炎，第 12 卷，第 1 期，页码 14，2020 年。[19]I. Tomonaga、H. Koseki、C. Imai、T. Shida、Y. Nishiyama、D. Yoshida、S. Yokoo 和 M. Os-aki，“2005 至 2019 年长崎县吸入性肺炎的发病率及特征”，《*BMC*肺部医学》，第 24 卷，第 191 期，2024 年。