Construction of an automatic score for the evaluation of speech disorders among patients treated for a cancer of the oral cavity or the oropharynx: The Carcinologic Speech Severity Index

Virginie Woisard, Mathieu Balaguer, Corinne Fredouille, Jérôme Farinas, Alain Ghio, Muriel Lalain, Michèle Puech, Corine Astesano, Julien Pinquier, Benoît Lepage

Received: 25 October 2020 | Revised: 15 August 2021 | Accepted: 5 October 2021

DOI: 10.1002/hed.26903

ORIGINAL ARTICLE

WILEY

Construction of an automatic score for the evaluation of speech disorders among patients treated for a cancer of the oral cavity or the oropharynx: The Carcinologic Speech Severity Index

Virginie Woisard MD, PhD[1,2,3] | **Mathieu Balaguer MSc**[1,4] | **Corinne Fredouille PhD**[5] | **Jérôme Farinas PhD**[2] | **Alain Ghio PhD**[6] | **Muriel Lalain PhD**[6] | **Michèle Puech**[1,2] | **Corine Astesano PhD**[3] | **Julien Pinquier PhD**[2] | **Benoît Lepage PhD**[1,7]

[1]ENT Department, University Hospital of Toulouse, Toulouse, France

[2]Oncorehabilation Unit, University Institute of Cancer of Toulouse Oncopole, Toulouse, France

[3]Laboratoire Octogone-Lordat, Jean Jaures University Toulouse II, Toulouse, France

[4]Institut de Recherche en Informatique de Toulouse, CNRS, Paul Sabatier University Toulouse III, Toulouse, France

[5]Laboratoire d'Informatique d'Avignon, Avignon University, Avignon, France

[6]Laboratoire Parole et Langage, Aix-Marseille University, Marseille, France

[7]USMR, Université Paul Sabatier Toulouse III, Toulouse, France

Correspondence
Virginie Woisard, ENT Department, Larrey Hospital, TSA 30030 University Hospital of Toulouse, 31059 Toulouse Cedex 09, France.
Email: woisard.v@chu-toulouse.fr

Funding information
Institut National Du Cancer, Grant/Award Number: 2014-1; French National Research Agency, Grant/Award Number: ANR-18-CE45-0008

Section Editor: Katherine Arnold Hutcheson

Abstract

Background: Speech disorders impact quality of life for patients treated with oral cavity and oropharynx cancers. However, there is a lack of uniform and applicable methods for measuring the impact on speech production after treatment in this tumor location.

Objective: The objective of this work is to (1) model an automatic severity index of speech applicable in clinical practice, that is equivalent or superior to a severity score obtained by human listeners, via several acoustics parameters extracted (a) directly from speech signal and (b) resulting from speech processing and (2) derive an automatic speech intelligibility classification (i.e., mild, moderate, severe) to predict speech disability and handicap by combining the listener comprehension score with self-reported quality of life related to speech.

Methods: Eighty-seven patients treated for cancer of the oral cavity or the oropharynx and 35 controls performed different tasks of speech production and completed questionnaires on speech-related quality of life. The audio recordings were then evaluated by human perception and automatic speech processing. Then, a score was developed through a classic logistic regression model allowing description of the severity of patients' speech disorders.

Results: Among the group of parameters subject to extraction from automatic processing of the speech signal, six were retained, producing a correlation at 0.87 with the perceptual reference score, 0.77 with the comprehension score, and 0.5 with speech-related quality of life.

The parameters that contributed the most are based on automatic speech recognition systems. These are mainly the automatic average normalized likelihood score on a text reading task and the score of cumulative rankings on pseudowords. The reduced automatic YC2SI is modeled in this way: $Y_{C2SIp} = 11.48726 + (1.52926 \times X_{averaged\ normalized\ likelihood\ reading}) + (-1.94e\text{-}06 \times X_{score\ of\ cumulative\ ranks\ pseudowords})$.

Conclusion: Automatic processing of speech makes it possible to arrive at valid, reliable, and reproducible parameters able to serve as references in the framework of follow-up of patients treated for cancer of the oral cavity or the oropharynx.

KEYWORDS
intelligibility, oral cancer, oropharyngeal cancer, severity index, speech disorders

1 | INTRODUCTION

The decrease in cancer mortality makes attention to the quality of life (QoL) after cancer a priority. This particularly concerns cancers of the upper aerodigestive tract (UADT), as their treatment can be especially mutilating.

However, there is a lack of uniform methods for the evaluation of functional results. Such tools, by measuring the impact after treatment of a given tumor localization on one or more of the most altered functions, would make it possible:

1. to complete the expression of treatment results by indexes of functional prognosis,
2. to better adjust treatment procedures to reduce their functional consequences.

For UADT cancers, it is mainly a question of the impacts on the major functions of communication (oral) and nutrition (swallowing).[1]

With regard to speech, the impact of the size of the tumor,[2,3] the localization of the tumor,[4] and the association of surgery with radiotherapy,[5] as well as the role of age[6,7] have been demonstrated. Patients' QoL will thus be compromised,[1] with particular consequences for social relations.[8] Evaluation of speech is therefore, essential, given the functional and social impact of its disorders.

Protocols have been proposed to evaluate voice disorders in cancers of the larynx. That is not the case for cancers of the oral cavity and the pharynx, which, while more frequent, have little impact on the voice, but especially affect speech articulation.[9]

Very few tools are available for the evaluation of speech disorders, although it is the most common symptom in cancers of the oral cavity and the pharynx.[10] This evaluation is based on perceptual evaluations, mostly conducted by speech–language pathologists (SLP). Speech is evaluated at the level of phonemic production in terms of acoustic-phonetic decoding (intelligibility), but also at the level of discourse (or "running speech intelligibility"), which involves, in the listener, cognitive mechanisms of mental compensation for the altered speech.[11] Moreover, these perceptual evaluation tools offer a very moderate interjudge and intrajudge fidelity.[12] The degree of familiarity of the listener with the speaker, or with the proposed task may improve predictability and thus the severity perceived by the judge. In addition, the emotional context or the mental availability of the judge at the time of the perceptual evaluation may modify the result.[13]

The field of automatic speech processing field refers to computer-based techniques capable of using linguistic and paralinguistic information from speech signals automatically in order to synthesize speech (text-to-speech tools), to recognize speech (speech-to-text tools like automatic speech recognition [ASR] systems), and analyze speaker-dependent information such as his/her identity, native language, regional accents, emotional state, or voice and speech impairments. Recently, the use of automatic speech processing tools has been recognized as an objective, standardizable method of evaluating communication disorders secondary to speech production disorders.[3,14] One particular German team of researchers and physicians at the University of Erlangen-Nuremberg[3,15,16] are working on the automatic quantification of speech intelligibility using an ASR system applied to UADT cancers. In 2008, Windrich et al.[15] first used this system for patients treated for oral cavity cancers and produced a rate of word recognition that was strongly correlated to a perceptual evaluation of intelligibility done by an expert jury ($r = -0.93$; $p < 0.01$).[16] In 2014, Middag et al.[17] presented a new method[18] that gave robust predictions of speech intelligibility when faced with changes of text and variations of accents in Flemish, which was applicable to patients treated for UADT cancer. This showed that automatic methods based on phonetic and phonologic alignment of speech allow a correlation between perceptual scores and automatic scores on the order of 0.80, for a general learning model. While the evaluation of speech by a jury of listeners remains the gold standard, the idea of calculating of an automatic severity index is gaining interest, and its development is accelerated thanks to the recent progress of automatic speech processing tools in machine learning (artificial intelligence).

However, none of these methods has led to the development of a tool that can be used in current practice. Besides the problems inherent in launching innovative solutions, the clinical interpretation of these measures

remains poorly developed. This problem refers to the question of measuring the functional deficit in the field of speech disorders.[19] Indeed, if the measurement of speech intelligibility is recognized as the measurement of the degree of alteration of speech production, it lacks correlation with measures of speech comprehension and accuracy in relation to prediction of people's communication abilities. Therefore, in order to estimate the severity of the functional deficit, the clinical interpretation of a speech severity score requires a real examination of issues addressing the consequences on the communicative capacities of individuals.

Thus, we propose the hypothesis that automatic speech processing allows the production of objective measures of the severity of speech pathologies in UADT oncology, making it possible to describe treatment results of protocols and supplementing survival rates.

The objective of this work is to (1) model an automatic severity index of speech applicable in clinical practice, that is equivalent to or superior to a severity score obtained by human listeners, via several acoustics parameters extracted (a) directly from speech signal and (b) resulting from speech processing and (2) derive an automatic speech intelligibility classification (i.e., mild, moderate, severe) to predict speech disability and handicap by combining the listener comprehension score with self-reported QoL related to speech.

2 | MATERIALS AND METHODS

This is a cross-sectional study based on audio recordings of patients with cancers of the oral cavity and the oropharynx conducted during a single session. The principle is:

1. to evaluate the alteration of speech production using the current reference method based on a perceptual task, that is, by a score of severity or of alteration of perceptual intelligibility as determined by a jury of listeners;
2. then, to participate in automatic methods of signal processing through different recordings of speech by the same subjects to find the best possible modeling of the reference score.

The automatic score thus obtained will then be studied in relation to other aspects of speech, that is, comprehension, oral communication, and QoL relative to speech handicap, in order to explore its capacity to describe speech disorders in terms of severity.

All of the data collection and recordings occurred at the same place for all speakers and on the same day for each speaker, during a follow-up consultation or a day hospitalization at the Institut Universitaire du Cancer de Toulouse (Toulouse University Cancer Institute) within the framework of the Carcinologic Speech Severity Index (C2SI) project, financed by the Institut National du Cancer (Grant INCa SHS No. 2015-135). Each subject had been previously informed about the nature of the C2SI project and the terms of participation in view of obtaining informed consent (via a consent form). To guarantee anonymity, a code was assigned to each subject. A declaration was made concerning data processing to the French national data protection authority (Commission nationale de l'informatique et des libertés) (number 1876994v 0 July 24, 2015), and a favorable decision was obtained from the research ethics committee of Toulouse Hospitals on May 17, 2016. The corpus established with the recordings are described in a preceding publication.[20] The methodologies applied for the analyses of this corpus are already published but are integral to this work.

2.1 | Speech recordings

Different types of production tasks were necessary in order to produce a speech sample that was sufficiently representative for finding a valid automatic score, for knowing which tasks will finally be adequate for use in clinical practice, and for specifying the significance of this score for the communication function.

There were five linguistically varied speech production tasks:

1. A task for producing a sustained vowel (/a/) that allowed the extraction of different acoustic parameters such as frequency, intensity, stability, noise, and so forth.
2. The production of 50 bisyllabic pseudowords (words that do not exist in French but follow its phonotactic rules) chosen randomly for each subject among the 89 346 possible combinations were created. The purpose of using these pseudowords was to limit the effects of context related to access to word meanings, and to explore speech quality at the acoustic-phonetic level, which enabled a score for each speaker to be computed that reflects the alteration of phonetic features per phoneme.[21]
3. A text reading task (first paragraph of "La chèvre de Monsieur Seguin" by Alphonse Daudet[22]), which enabled an overall subjective measurement of severity and of intelligibility.[23]
4. A task of describing an image randomly chosen from a set of 10 images on the same theme,[24] which also

enabled an overall subjective measurement of severity and of intelligibility. Contrary to the reading task, however, this descriptive task left the speaker some freedom of expression, while providing a general context and forcing the use of a given lexical field.

5. A reading task consisting of a list of 50 short affirmative sentences, 25 containing true information, and 25 containing false information. These 50 sentences make up part of a corpus of 300 phrases, that is, 150 pairs of true/false sentences, of which the true or false nature can only be identified by the last lexical unit (e.g., "Hens lay eggs" vs. "Hens lay fruits"). They enable the evaluation of comprehensibility by a task called the "sentence verification task".[25,26]

Finally, three of the tasks were used for automatic processing, that is, production of the vowel /a/, production of 50 pseudowords and the reading (first paragraph of "La chèvre de Monsieur Seguin"), and two tasks were used for perceptual analysis, that is, the description of an image resulting in the severity reference score, and the reading of a list of 50 short affirmative sentences resulting in a comprehensibility score.

2.2 | Population (speakers)

2.2.1 | Inclusion criteria

All the patients seen in consultation posttreatment for cancer of the oral cavity or the oropharynx between 2015 and 2017 at the Institut Universitaire du Cancer Oncopole de Toulouse who met the following criteria were included: being in a "chronic" phase (i.e., having completed the treatment protocol at least 6 months previously) and being in clinical remission, to ensure that the speech disorder would be as stable as possible. Not included were patients presenting a speech disorder potentially related to another pathology. A control group of 35 subjects was also recruited, composed of accompanying persons of the patients, to favor recruitment close in age, lifestyle, and location.

2.2.2 | Clinical data collection

Demographic information and clinical and treatment data for the patients were also collected including the anatomic area affected by the cancer; the TNM classification of the tumor[27]; type of treatment: (tumor surgery, lymph node surgery, radiotherapy, and/or chemotherapy); and the time since the end of treatment (in months).

2.2.3 | QoL relative to speech questionnaire

All the participants were asked to fill out questionnaires on QoL related to speech: the Speech Handicap Index (SHI)[28] and the Phonation Handicap Index (PHI).[29] These two questionnaires have been the subject of a study[30] showing their equivalence. They are composed of 30 questions for the SHI and 15 for the PHI, each scored on a scale of the Likert type with 5 levels (0 never, 1 almost never, 2 sometimes, 3 almost always, and 4 always), for a score of 0–120, and 0–60, respectively. The PHI offers the advantage for our objective of taking into account the functional dimension of oral communication, thanks to a "communication" field including the following questions: "I speak with friends and neighbors or relatives less often because of my speech," "I have difficulties to express orally what I need (drink, eat, go to the restroom...)," "I am hindered from expressing my thoughts, my opinions," "I have trouble communicating with unfamiliar people," "I am asked to repeat because of my speech difficulties." It also includes a symptom field and a psychosocial field following the same model, that is, five questions for a subscore of 0–20 points per field.

Finally, we kept the PHI for assessing the impact of speech alteration and especially the functional dimension of the impact on oral communication.

2.3 | Perceptual analysis of speech

Several listener juries were necessary to carry out this work. The choice of the jury composition was determined by our objectives and the feasibility related to the quantity of the recordings. As the quality of the reference score was crucial for modeling, we chose a task for trained professionals measuring the intelligibility and perceptual severity. Moreover, we wanted to model the deficit and to assess the impact of this deficit in the real life of the patient. For this reason, we chose an expert jury for the reference score, a naïve jury for the comprehensibility task and a self-questionnaire for the QoL relative to speech.

2.3.1 | Perceptual severity reference score

An expert jury was established to obtain the perceptual reference score. It was composed of six SLP and phoniatric experts in evaluating patients with speech disorders. The raters had at least 7 years of experience. Because of our previous work on the difference between the perception of the degree of intelligibility and the degree of severity,[19] the determination of perceptual

scores of severity and intelligibility was achieved using the reading task and the image description task.

The experts were given instructions: (1) to first listen to the stimuli and to score each one based on the quality of voice, resonance (nasal and pharyngeal), prosody, and phonemic production, using a 4-point ordinal scale (0–3, 0 no alteration, 3 significant alteration), and (2) to afterward use an visual analog scale for determining the level of severity and alteration of intelligibility. Alteration of intelligibility was defined as "the degree to which the message of the speaker can be understood by a listener[31] in terms of proportion of speech".[32] Severity was defined as "the quality of the emission of the acoustic-phonetic code,[33] including the measurement of the flow of speech,[34] and other temporal and/or prosodic parameters in relation with the discomfort felt.[35]" For that, the stimuli may be listened to without any time limitation. Both scores range from 0 to 10 (0 for important alteration, 10 for no alteration).

The results of this trial were published,[36] and the outcomes produced by the instructions about the degree of severity of the image description task were retained as a reference for perceptual judgment. This task provided the best satisfactory inter-rater agreement (intraclass correlation coefficient of 0.69) and the best distribution of the scores.

2.3.2 | Comprehensibility score

For the score of comprehensibility obtained using the "sentence verification task," a cohort of 146 naïve listeners (i.e., not accustomed to the altered speech to avoid the effect of high level processes on their perception of speech)[37] performed a perceptual evaluation of the speakers. Each participant had to verify the veracity of 100 sentences. One score per speaker was calculated which was equivalent to the average of each of the perceptual evaluations received during the test. Since each item was evaluated three times by the listeners, this score lies between 0 and 3. The method used is presented in the publication of Nocaudie et al..[26]

2.4 | Automatic processing of recordings

2.4.1 | Extraction of acoustic parameters

Three software programs were used for acoustic processing. The software openSMILE 3.0[38] permits the extraction of over 1000 acoustic parameters on a sustained vowel. PRAAT 6.1.16[39] was used for the measurements of formants, and VOCALAB 4[40] to obtain composite criteria on the task of /a/ production and on text reading.

After the acoustic processing was applied, a theoretical choice, which was a prerequisite for the analysis, was applied to the significant differences of scores between patients and control subjects and/or to the fact that the parameters would be used in clinical practice, given their very great number.[40]

The main parameters studied are listed in Table 1.

2.4.2 | Speech processing

Different studies involving automatic speech processing had been conducted on a first batch of recordings before being applied to the whole corpus. Among these studies, we can report the use of an i-vectors-based approach derived from the automatic speaker recognition field (automatically recognizes the identity of a speaker based on his/her voice) for predicting a speech intelligibility score[41]; the use of an automatic system for detecting speech abnormalities, especially designed for analyzing speech impairments[42]; and the proposal of a "robotic" listener for automatically performing the phonetic-acoustic decoding of the pseudoword productions and for providing an automatic measure of speech intelligibility.[43] Automatic speech alignment was involved in the last two studies, which consists of aligning the sequence of expected phonemes (corresponding to a word or a pseudoword pronounced by a speaker in this case) on the corresponding speech signal. The outputs of the automatic alignment are a set of start and end time boundaries in the speech signal for the sequence of expected phonemes. In a second step, all possible phonemes are placed in competition on each available phonetic segment derived from the automatic alignment in order to obtain a sequence of phonemes that is closest to what has been produced and not to what is to be produced. This approach automatically computes a normalized likelihood score. The higher this score, the more the sequence of expected phonemes is considered to have been correctly pronounced by the recorded speaker.

Finally, we investigated the use of an ASR system for recognizing each pseudoword pronounced among the 86 346 occurrences available in a phonological dictionary involved in the ASR system. Due to the great acoustic confusion between pseudowords, we did not consider the pseudoword automatically recognized by the ASR system, as commonly done, but its recognition ranking among the 86 346 occurrences. A ranking of 1 indicates that the target pseudoword was classified as the best occurrence recognized by the ASR system whereas a ranking of 86 346 is considered as the worst occurrence recognized. By adding the rankings of all the pseudowords produced by the subject, we end up with the calculation of a score, denoted as the "cumulative rank score."

By comparing all these approaches based on automatic speech processing following our objective, four scores were

TABLE 1 Mean and SD for the patients and control subjects obtained from acoustic processing. *p*-Value of the difference of means obtained

	Patients		Controls		
	Mean	**SD**	**Mean**	**SD**	***p*-Value**
On the vowel /a/					
Duration (s)	5.06	3.49	9.82	6.57	**0.008**
Fo median	153	43	176	53	**0.025**
Fo interquartile range	13	22	8	15	**0.028**
Jitter local median (Hz)	0.0055	0.0035	0.0038	0.0026	**0.001**
Jitter local interquartile range (Hz)	0.0061	0.0046	0.0038	0.0027	**0.001**
Height instability (Hz)[a]	2.12	2.68	0.94	0.84	**0.001**
Amplitude instability (dB)[b]	1.04	0.47	0.80	0.29	0.072
Ratio signal/noise (%)	1.91	1.69	1.11	0.76	**0.005**
Log-HNR median (dB)	10.09	31.93	15.76	23.79	0.405
Quality of the attack[c]	1.73	1.33	1.10	0.55	**0.005**
Harmonic poverty[d]	2.10	1.20	1.36	0.81	**0.009**
Mel-frequency cepstral coefficients	2.79	0.39	4.67	0.63	**0.047**
On text reading					
Duration (s)[e]	32.78	9.57	23.76	4.86	**<0.0001**
Range of Fo (1/2 tones)	9.43	3.31	11.29	2.79	**0.023**
Height instability (Hz)[a]	9.67	3.20	13.31	4.53	**0.029**
Frequency distance /m/−/s/ (1/2 tones)[f]	15.93	10.38	24.14	11.50	**0.002**
Harmonic poverty	4.97	1.37	6.18	0.76	**0.048**

Note: All bold values are significant ($p < 0.05$).

[a]Instability of vocal height is calculated from jitter at medium, long, and very long term.

[b]Instability of amplitude depends on the variation of the amplitude of the signal over time (or shimmer) at medium, long, and very long term.

[c]Quality of attack of the sound is measured by analysis of the height and amplitude instability and the signal/noise ratio during the first 300 ms.

[d]Harmonic poverty is based on the medium to long term of the spectrum and corresponds to the number of harmonics included between the frequencies 75 and 2500 Hz.

[e]The gross duration of the reading duration was automatically collected as a parameter dependent on speech flow.

[f]Frequency distance /m/−/s corresponds to the distance between the formant peaks of the consonants /s/ and /m/ in the "Monsieur Seguin" sequence.

retained from the task of pseudoword production task: the score of cumulative ranks: (1) the higher the score, the more severe the speech disorder; (2) the average rate of detection of abnormalities in the speech signal associated with pseudowords pronounced by the speaker; the higher the score, the more severe the speech disorder; (3) the average normalized likelihood score derived from the approach based on automatic speech alignment reported above: the lower the score, the more severe the disorder; and (4) the number of "features" of average deviation per phoneme obtained from an automatic decoding issued from the "robotic" listener mentioned above. A "feature" is a phonetic characteristic of a sound of the French language: there are five for the vowels (nasal, back, rounded, high, and open), and six for the consonants (vocalic, continuous, nasal, voiced, compact, and acute). For example, there is one feature of deviation between the [a] and [an]—the nasal feature; two features between the [p] and [d]—the voiced and the acute features. Thus, the greater the number of features of average deviation per phoneme, the more the perception deviates from the expected form and therefore, the more serious the disorder.[44]

In the same way as for the pseudowords production, automatic analysis derived from the approach based on automatic speech alignment leads to the calculation of an average normalized likelihood score on the text reading.

The parameters are presented in Table 2. They show that automatic speech processing also allows for distinction between patients and controls.

2.5 | Automatic severity score

2.5.1 | Modeling of the perceptual severity score

The modeling was done with a predictive approach using a classic linear regression model. We chose perceptual

TABLE 2 Results of scores obtained from automatic processing of speech with the *p* value of the difference of means obtained

	Patients		Controls		
	Mean	**SD**	**Mean**	**SD**	***p*-Value**
On the production of pseudowords					
Score of cumulative ranks	210 516	250 446	16 908	31 603	<0.0001
Averaged rate of detection of abnormalities	43.19	14.09	25.54	14.42	<0.0001
Number of features of average deviation per phoneme	1.28	0.63	0.48	0.22	<0.0001
Averaged normalized likelihood score	−4.04	0.94	−3.16	0.93	<0.0001
On text reading					
Averaged normalized likelihood score	−3.11	1.00	−1.76	0.43	<0.0001

severity as the dependent variable, that is, the score defined above in Section 2.3.1, and the explanatory variables were the acoustic and automatic voice and speech parameters retained following the validity study. After having eliminated certain redundant variables to avoid phenomena of overadjustment related to a potential colinearity between variables, the pertinence of the selection of the parameters retained was verified by a confirmatory factor analysis, and a measurement of the internal consistency of the items retained (Cronbach's alpha). For that, transformations of variables allowed normalization of parameters. Moreover, verification that conditions of application of the model had been respected was achieved by control of equality of variances (Breusch–Pagan test) and the distribution of residuals (Shapiro–Wilk test). Finally, the quality of the measurement by the modeled C2SI score was evaluated by the calculation of Spearman's correlation coefficient with the perceptual severity score and by a Bland and Altman plot representing the difference between the C2SI scores on the one hand, and perceptual severity on the other hand, in relation to the average of these two scores.

All statistical analyses were performed using the software Stata 14.2.[45]

Initially, descriptive univariate analyses were conducted. For quantitative variables, a test of normal distribution (Shapiro–Wilk test) was performed and allowed the choice of parametric or nonparametric tests for bivariate analyses.

In order to compare the different means/medians obtained for our variables, the parametric Student's test, or the nonparametric Mann–Whitney *U* test for binary variables, and the Kruskal–Wallis test for variables of several categories, were used.

2.5.2 | Validity study

The validity study includes an analysis of the validity of construct and of criterion.

The analysis of the validity of construct was conducted on two points:

- An analysis of converging and discriminating validities by a matrix of Spearman or Pearson correlation coefficients. The interpretation of the matrix follows the recommendations of Mukaka et al..[46] The correction of Bonferroni was applied for the interpretation of the statistical significance of the results, in order to remedy the bias linked to the performance of multiple tests.
- An analysis of the results obtained for the extreme groups (patients and controls); the scores obtained by the controls should be significantly different from those obtained by the patients overall in the set of parameters for voice and speech signal (clinical validity).

To evaluate the validity of criterion, we compared the different results with our reference test (perceptual severity score during the image description test).

Once this modeling terminated, a score that would give the same result with less speech production by the patients, that is, a more "economical" version, was sought by proceeding with a descending method.

2.5.3 | Description of the population according to the severity of speech production with the automatic C2SI score

To determine the clinical meaning of the speech severity score, the data obtained on healthy subjects were used to determine the diagnostic cut-off points between the control subjects and patients with speech disorders. Normality thresholds were calculated from the mean of the scores and the mean SD of control subjects when the data distribution followed the normal law, either according to increase or decrease of the score for the normality: mean $\pm$ (2 $\times$ SD).[47] If not, the threshold will be estimated from the distribution of the quantiles at the 95th percentile.

Then, the observation of the distribution of the C2SI score and the calculation of the receiver operating characteristic (ROC) curves in relation to comprehension and the communication field of the PHI, allowed determination of the scores of C2SI associated with the best probability that the speech disorder would have consequences on social life through oral communication. For that, the C2SI score corresponding to a sensitivity of 90% was retained.

A description of the speech production disorders of our population could thus be proposed in three classes of severity included between the maximum C2SI score, the normality threshold of C2SI, the severity threshold determined by the impact on comprehension, and the communication field of the PHI.

3 | RESULTS

3.1 | Population

Here, 87 patients treated for cancer of the oral cavity or the pharynx were registered (51 men and 36 women), average age 65.8 years (range 36–87 years). Of these, 40.2% had cancer of the oral cavity, while 59.8% had an oropharyngeal tumor (Table 3).

For the oral cavity, the origin of the tumor was mostly the floor of the mouth,[15] followed by the tongue,[8] the retromolar region,[6] and the mandible.[5] For the oropharynx, it involved especially the tonsils,[25] the base of the tongue,[13] and the soft palate.[4] For nine tumors of the oropharynx, their extension prevented attribution of a precise starting point.

With regard to tumor classification, none of the patients had metastasis. For the classification, 12.64% were T1, 37.93% T2, 13.79% T3, and 35.66% T4. Then, 31% of patients were N0, 25% N1, 37% N2 a, b, or c, and 7% N3.

Also, 89% of patients had undergone surgical treatment. This treatment was followed by chemoradiotherapy in 45% of cases, or by radiotherapy alone in 39% of cases. No additional treatment was given in 5% of cases. A lymph node dissection was associated with removal of the tumor in 80% of cases. Eleven percent of patients were treated by radiotherapy or radiochemotherapy alone. Radiation involved the nodal area in 60% of cases. The various combinations are shown in Table 3. Also, 66% of the patients underwent surgical reconstruction. No analysis was made about the type of reconstruction because of the large quantity of missing data in that field, with 15 different types of reconstruction and just a few subjects in each.

The posttreatment period during the recordings was 64.2 months on average (range of values 6–239 months) with a median of 39 months.

3.2 | Perceptual judgment

The results of the perceptual severity score and the comprehension score are shown in Figure 1.

For the severity score obtained from the image description task, that is, our reference score, the mean is 6.06 (±2.36) for patients and 9.63 (±0.41) for controls. The description of this score in relation to the oncological characteristics of the patients was published previously.[36] We found some voice abnormalities even though they were much less significant than the speech articulation abnormalities. This observation may be related to the significant difference between patients and controls regarding the scores of the acoustic parameters and the fact that patients were treated with cervical irradiation for metastatic nodes.

The comprehension score is a mean of 2.39 (±0.63) for patients and 2.88 (±0.05) for controls for a maximum score of 3, which corresponds to perfect comprehensibility.

3.3 | Results of the self-administered questionnaire on speech handicap

The scores of the PHI (Figure 2) are, respectively, for the patients and controls for the Symptom (S) field, 8.27 (±4.64) and 1.21 (±1.69); for the Communication (C) field, 6.13 (±4.53) and 0.23 (±0.7); for the Psychosocial (E) field, 6.7 (±5.19) and 0.28 (±0.9). The total score is 21.11 (±13.35) for patients and 1.75 (±2.56) for controls.

3.4 | Determination of the automatic severity score

3.4.1 | Study of the validity of parameters

Validity of construct

Analysis of convergent and discriminant validity clearly shows two groups of parameters (Table 4). The first consists of acoustic parameters that can be described as vocal (orange insert), in contrast to the second group composed of speech parameters obtained from automatic processing of speech and from the duration of reading that represents a flow of read speech (blue insert).

The voice parameters are correlated between each other with Spearman or Pearson correlation coefficients superior to $|r| > 0.30$, except for correlations mainly linked to the quality of attack and harmonic weakness.

Similarly, the scores obtained from speech parameters evaluated on pseudoword production, but also the average normalized likelihood score issued from the reading task are all very highly correlated (>0.69). Only the duration of speech is weakly correlated with the likelihood scores.

TABLE 3 Treatment offered to patients

Treatment combination				*n*	Surgical procedure of tumor site	*n*
TS	NS	Ch	RT		Limited resection (tonsil or soft palate or mouth floor)	8
X	X	X	X	35	Partial glossectomy	10
X		X	X	3	Total glossectomy	2
X	X		X	30	Pelvi ± glosso ± mandibulectomy	27
X	X	X		1	Oropharyngectomy	26
X	X			4		
	X	X	X	2		
	X		X	4		
		X	X	7		
			X	1		

Note: On the right side, treatment combination. On the left side, surgical procedures.
Abbreviations: Ch, chemotherapy; NS, node surgery; RT, radiotherapy; TS, tumor surgery.

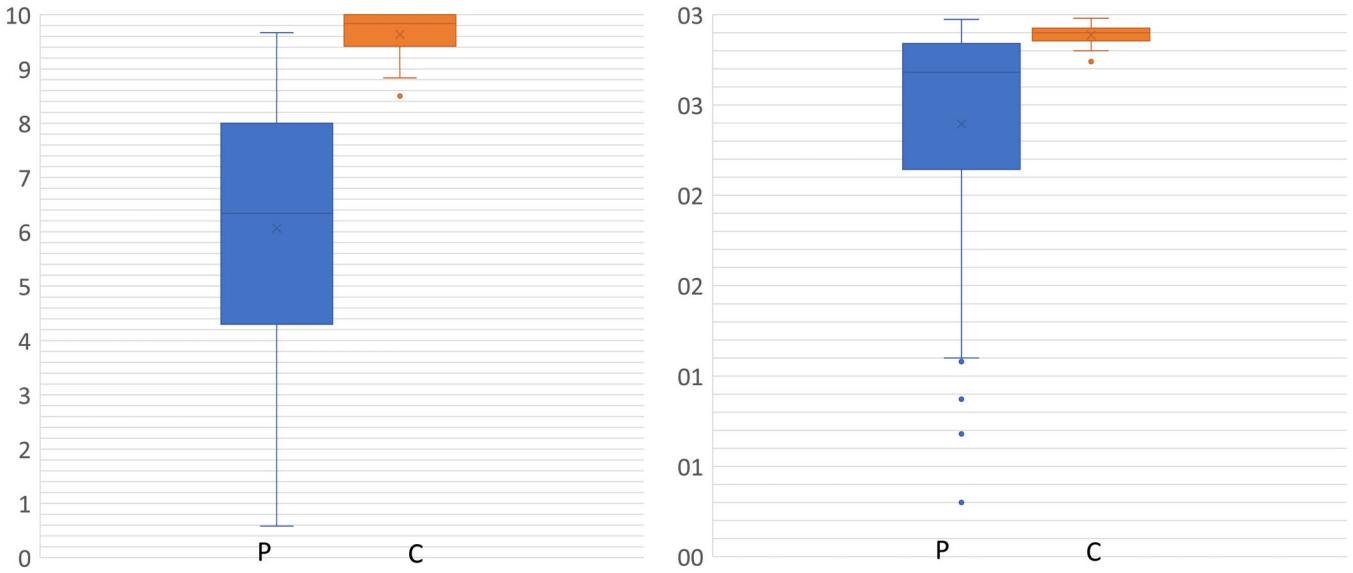

FIGURE 1 Distribution of severity scores at the left (scores range from 0 to 10; the lower the score, the greater the severity), and comprehension scores at the right (scores range from 0 to 3; the lower the score, the greater the severity), P for patients, C for controls [Color figure can be viewed at wileyonlinelibrary.com]

Only two weak correlations are found between the frequency parameters of voice and the score of cumulative rank parameters (Table 4).

The analysis of results obtained from the extreme groups patients/controls (Tables 1 and 2) shows significant differences ($p < 0.05$). Only two tests failed to discriminate between subjects and controls: instability of amplitude and the mean of log-HNR. These measures are voice related.

Criterion validity

The analysis of criterion validity includes the study of correlations between the different parameters of voice and speech, the perceptual severity score (our gold standard), and the speech handicap (PHI) (Table 5). It shows that the scores obtained from the pseudoword production task, as well as the likelihood score on reading (i.e., speech scores) are all strongly correlated with the perceptual severity score ($|r| > 0.62$). They are also correlated, more or less strongly, with the speech handicap scores. Notably, the latter show stronger correlations for the "Communication and social handicap" field than for the "Symptom" field of the PHI. As for the voice tasks, they are weakly correlated with the questionnaire scores ($|r| < 0.18$), and only the interquartile deviation of the fundamental frequency and height instability show a correlation coefficient superior to 0.25 with the perceptual severity score.

For the purpose of construction of our score therefore, the most pertinent automatic parameters were

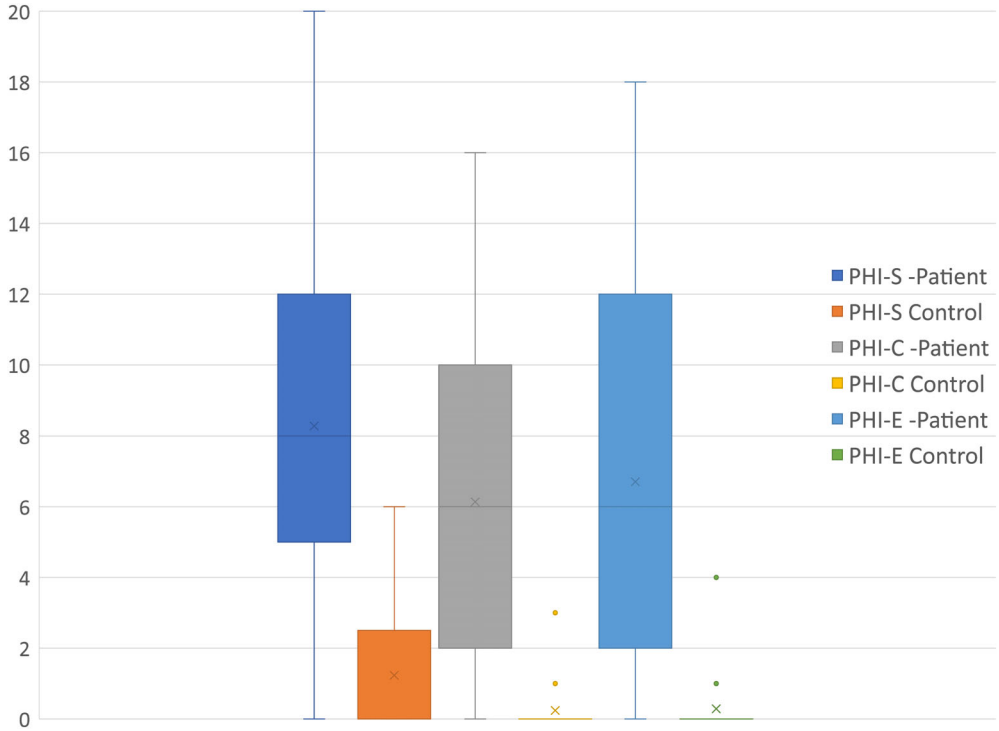

FIGURE 2 Distribution of scores of the Phonation Handicap Index [Color figure can be viewed at wileyonlinelibrary.com]

selected, meaning those that respected the conditions of validity.

In the end, we selected (1) two voice parameters: the interquartile deviation of fundamental frequency and height instability (both being measured by the production of the sustained /a/); and (2) five speech parameters, four of which were obtained from the production of pseudowords (automatic mean rate of deviation per phoneme, average normalized likelihood score, score of cumulative ranks, and average rate of detection of abnormalities), as well as the likelihood score obtained from the reading task.

Verification of the pertinence of the selection was done with a confirmatory factor analysis and a measurement of internal consistency.

The confirmatory factor analysis on the seven retained parameters verified that the two groups "voice" and "speech" were internally homogeneous, and clearly distinct from each other, even though there remained considerable variability, especially in the voice parameters (uniqueness at 0.65 and 0.68 for the voice group, and between 0.37 and 0.09 for the speech group).

For the internal consistency of the "speech" field, we calculated a Cronbach's alpha coefficient. The alpha coefficients were all very high (≥0.88) and the overall alpha of the field was 0.91.

For the "voice" field, comprising two parameters, Spearman's correlation gave a coefficient of 0.71.

Analysis of all these results led us to exclude:

- the variables of "mean rate of deviation per phoneme" and "automatic mean rate of deviation per phoneme" on the production of pseudowords compared with other speech variables (e.g. the correlation between this variable and the likelihood score on the production of pseudowords was very high, at 0.90), in order to avoid phenomena of overadjustment related to a potential colinearity between variables;
- the variable concerning duration of reading (equivalent to a measurement of flow of speech) because of the weakness of its correlations with the other parameters of the "speech" field.

3.4.2 | Modeling of the score

The automatic score was then constructed from six parameters: two from the "voice" field, and four from the "speech" field. Multivariate analysis was performed concerning the 59 patients with no missing data on the automatic parameters retained (Table 6).

The conditions of application of this model were respected, with an equality of variances (Breusch–Pagan test: $p = 0.655$) and normally distributed residuals (Shapiro–Wilk test: $p = 0.416$).

The test of overall adjustment of the model to the data was highly significant ($p < 0.001$) with a high coefficient of determination R^2 at 0.757.

Thus, our automatic score YC2SI based on the dependent variable "gold standard" of speech disorder severity was modeled in this way: $\text{YC2SI} = 11.26482 + -0.0049184 \times_{\text{XF0-IIQ}} + -0.0946604 \times_{\text{Xheight instability on /a/}} + -0.147016 \times_{\text{Xaverage normalized likelihood score pseudowords}}$

TABLE 4 Correlation matrix between the automatic results among the patients

	IQR of F0	Harmonic poverty	Quality of attack	Jitter median	Jitter IQR	Height instability	Amplitude instability	Median of log-HNR	Ratio signal/noise	Features of average deviation, Pseudowords	Averaged normalized likelihood, Pseudowords	Averaged normalized likelihood, Reading	Duration, Reading	Score of cumulative ranks, Pseudowords
Harmonic poverty	0.06	1.00												
Quality of attack	0.49	0.21	1.00											
Jitter median	0.39	0.54	0.21	1.00										
Jitter IQR	0.51	0.45	0.37	0.87	1.00									
Height instability	0.71	0.22	0.60	0.36	0.46	1.00								
Amplitude instability	0.49	0.38	0.35	0.47	0.49	0.57	1.00							
Median of log-HNR	−0.48	−0.40	−0.25	−0.59	−0.55	−0.57	−0.59	1.00						
Ratio signal/noise	0.32	0.58	0.33	0.39	0.43	0.42	0.43	−0.48	1.00					
Features of average deviation, Pseudowords	0.23	−0.05	0.09	0.04	0.07	0.24	0.10	−0.10	0.02	1.00				
Averaged normalized likelihood, Pseudowords	−0.27	0.13	−0.13	0.01	−0.05	−0.20	−0.06	0.05	0.06	−0.90	1.00			
Averaged normalized likelihood, Reading	−0.23	−0.04	−0.04	−0.20	−0.16	−0.11	−0.01	0.17	0.01	−0.70	0.76	1.00		
Duration, Reading	0.25	−0.08	0.15	−0.05	−0.03	0.22	−0.03	−0.03	0.19	0.28	−0.31	−0.39	1.00	
Score of cumulative ranks, Pseudowords	0.44	−0.05	0.13	0.05	0.11	0.38	0.10	−0.20	0.21	0.69	−0.72	−0.78	0.58	1.00
Averaged rate of detection of abnormalities, Pseudowords	0.15	−0.15	0.02	−0.01	0.09	0.06	−0.09	0.01	−0.08	0.76	−0.76	−0.72	0.23	0.69

Note: In dark gray, $r \geq 0.50$; in light gray, $0.30 \leq r < 0.50$.
Abbreviation: IQR, interquartile range.

TABLE 5 Matrix of correlation between automatic results, perceptual severity scores, and questionnaire scores

		Speech handicap			
	Perceptual gold standard	**PHI**	**PHI – F**	**PHI – C**	**PHI – E**
IQR of F0	−0.29	0.16	0.15	0.15	0.17
Harmonic poverty	0.05	−0.06	−0.09	−0.07	−0.05
Quality of attack	0.01	−0.12	−0.10	−0.10	−0.12
Jitter median	−0.14	0.06	0.05	0.11	0.04
Jitter IQR	−0.11	0.01	0.01	0.03	0.01
Instability of the height of /a/	−0.25	0.09	0.09	0.10	0.09
Amplitude instability	0.01	−0.04	0.00	0.00	−0.10
Median of log-HNR	0.23	−0.09	−0.04	−0.10	−0.11
Ratio signal/noise	−0.06	0.03	0.07	−0.05	0.05
Features of average deviation, Pseudowords	−0.62	0.29	0.23	0.37	0.21
Averaged normalized likelihood score, Pseudowords	0.68	−0.37	−0.28	−0.47	−0.30
Averaged normalized likelihood score, Reading	0.85	−0.47	−0.33	−0.55	−0.44
Duration, Reading	−0.51	0.47	0.36	0.48	0.45
Score of cumulative ranks, Pseudowords	−0.81	0.47	0.35	0.50	0.45
Averaged rate of detection of abnormalities, Pseudowords	−0.64	0.30	0.21	0.36	0.27

Note: In dark gray, $r \geq 0.50$; in gray, $0.30 \leq r < 0.50$; in light gray, $0.25 \leq |r| < 0.30$.
Abbreviation: IQR, interquartile range.

$+ 1.391981 \times$ Xaverage normalized likelihood score reading $+ -2.09\text{e-}06 \times$ Xscore of cumulative ranks pseudowords $+ (-0.0111486 \times$ Xaverage rate of detection of abnormalities pseudowords).

The quality of the measurement of our automatic C2SI score constructed by modeling for each of the patients in our sample was confirmed with a Spearman's correlation coefficient with the perceptual severity score at 0.87. The Bland and Altman plot representing the difference between the C2SI scores on the one hand, and perceptual severity scores on the other hand, in terms of the average of these two scores, is shown in Figure 3.

The analysis of this plot shows that 95% of the differences between the C2SI and severity scores are in the range ±2.16. We do not find systematic bias, although a slight tendency to underestimate the measurement for the high scores (higher than 8/10) may be noted. That concerns therefore, the patients with a well-preserved quality of speech. Moreover, Pitman's test of difference in variance shows that there is no bias related to the variation in deviations ($p = 0.037$).

Finally, it was possible to calculate a more economical version of this score, by following a descending procedure. Step by step, at each stage, the least significant variable was removed from the model. A model with two explanatory variables was then constructed (Table 6). This model also verifies all application conditions. The test of overall adjustment of the model to the data is highly significant ($p < 0.001$). The coefficient of determination R^2 is high (0.762). This reduced C2SI score presents the same properties as the complete score: its correlation with the perceptual severity score is 0.87, and the Pitman test between these variables does not show bias ($p = 0.022$).

The reduced automatic YC2SI score is modeled in this way: $Y_{C2SIp} = 11.48726 + (1.52926 \times$ Xaverage normalized likelihood reading$) + (-1.94\text{e-}06 \times$ Xscore of cumulative ranks pseudowords).

There are then, two versions of this C2SI score presenting the same overall performances in our population: a very good correlation with the reference score ($r = 0.87$ with the speech disorder severity score measured on the image description task).

For future results, we shall use the reduced version that will require the patient to read a list of pseudowords and a paragraph of the text of "La chèvre de M. Seguin."

3.5 | Description of the population according to the severity of the speech production disorder with the automatic C2SI score

The reduced C2SI score ranges from 1.6 to 11.47. It is on average 6.33 (±1.88) for patients and 8.87 ± 0.89 for controls (Figure 4), with a significant difference ($p < .001$). A

TABLE 6 Construction of the automatic C2SI and the reduced C2SI scores by multivariate analysis modeling

	Bivariate analysis		C2SI multivariate analysis			Reduced C2SI multivariate analysis		
Severity	**Coeff.**	***p*-Value**	**Coeff.**	**CI at 95%**	***p*-Value**	**Coeff.**	**CI at 95%**	***p*-Value**
Constant			11.26482	9.57; 12.96	**<0.001***	11.48726	10.50; 12.47	**<0.001***
Fundamental frequency – interquartile range	−0.29	**0.01***	−0.0049184	−0.02; 0.01	0.45			
Height instability on /a/	−0.25	**0.04***	−0.0946604	−0.29; 0.10	0.34			
Automatic likelihood score on logatomes	0.68	**<0.001***	−0.147016	−0.77; 0.48	0.64			
Automatic likelihood score on reading	0.85	**<0.001***	1.391981	0.84; 1.94	**<0.001***	1.52926	1.16; 1.90	**<0.001***
Accumulation of rows on logatomes	−0.81	**<0.001***	−2.09e-06	−3.86e-06; −3.07e-07	**0.02***	−1.94e-06	−3.36e-06; −5.12e-07	**0.008***
Abnormality rates on logatomes	−0.64	**<0.001***	−0.0111486	−0.05; 0.02	0.52			

Note: All bold values are significant ($p < 0.05$).
Abbreviation: C2SI, Carcinologic Speech Severity Index.

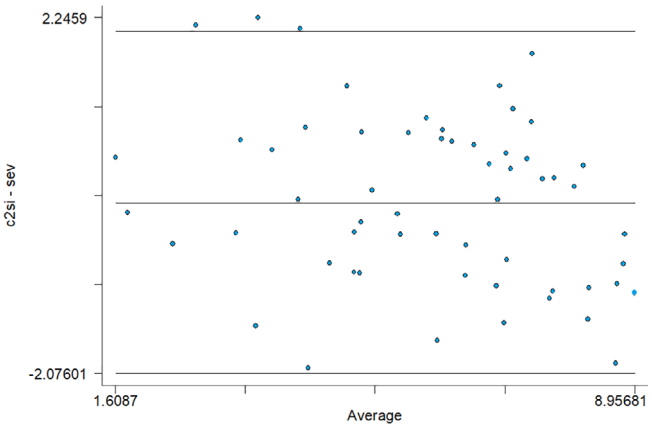

FIGURE 3 Bland and Altman plot between automatic Carcinologic Speech Severity Index (C2SI) score and perceptual severity score [Color figure can be viewed at wileyonlinelibrary.com]

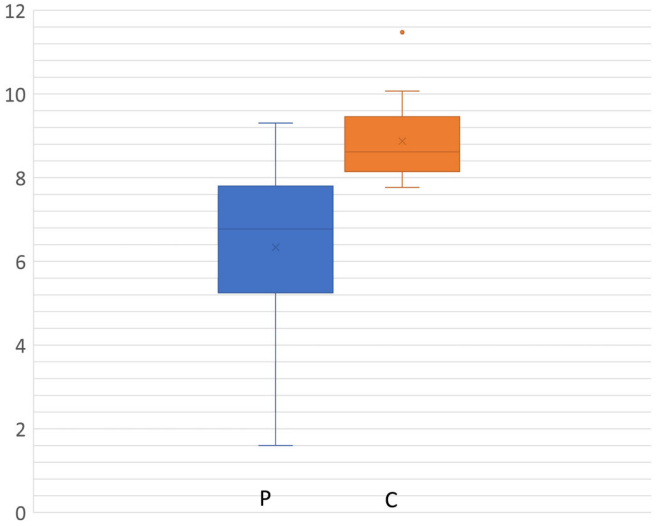

FIGURE 4 Boxplot of the Carcinologic Speech Severity Index (C2SI) score (P for patients, C for control subjects) [Color figure can be viewed at wileyonlinelibrary.com]

significant difference is also found in relation to the size of the tumor. Patients treated for tumors of small size (T1 or T2) have a significantly higher score than patients having presented a T3 or T4 tumor: respectively, 6.98 ± 1.71, and 5.58 ± 2.39 ($p = 0.001$). In contrast, the C2SI score does not allow demonstration of significant difference by anatomic region.

The calculation of the normality threshold from the normal population distribution is $8.87 - (2 \times 0.89)$ (mean-2DS), that is, 7.09.

The analysis of correlations between the C2SI score and the comprehension task is at 0.77 ($p < 0.001$). It is 0.5 ($p < 0.001$) for the PHI with a better correlation for the communication field at 0.53 ($p < 0.001$) (Figure 5 (B)), while it is 0.35 for the Symptom field ($p = 0.035$), and 0.45 for the psychosocial field ($p = 0.012$) (Figure 5).

By considering the distribution of C2SI scores, a first severity threshold can be estimated according to the theoretical normality threshold of the communication field of the PHI control population determined at the 95th percentile, which thus equals 1, and of the comprehension field, thus $2.88 - (2 \times 0.05) = 2.78$.

For these different values, the C2SI score for 90% sensitivity is ≥2.791 for comprehension, and ≥3.25 for the

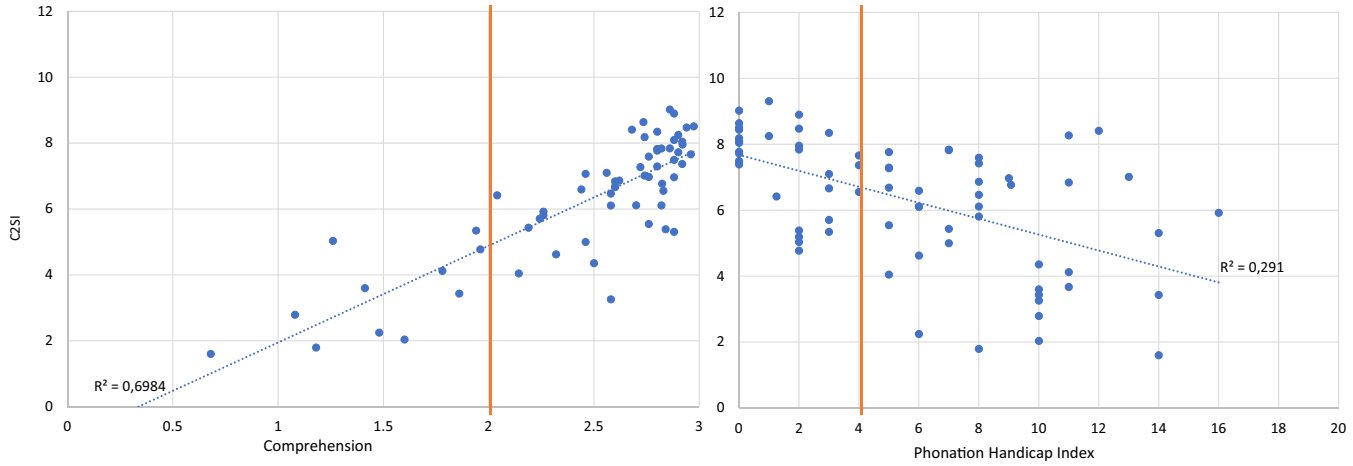

FIGURE 5 Scatterplot of the Carcinologic Speech Severity Index (C2SI) score according to the comprehensibility score and the communication field of the PHI. Dotted line: trend line. Continuous vertical line: severity threshold [Color figure can be viewed at wileyonlinelibrary.com]

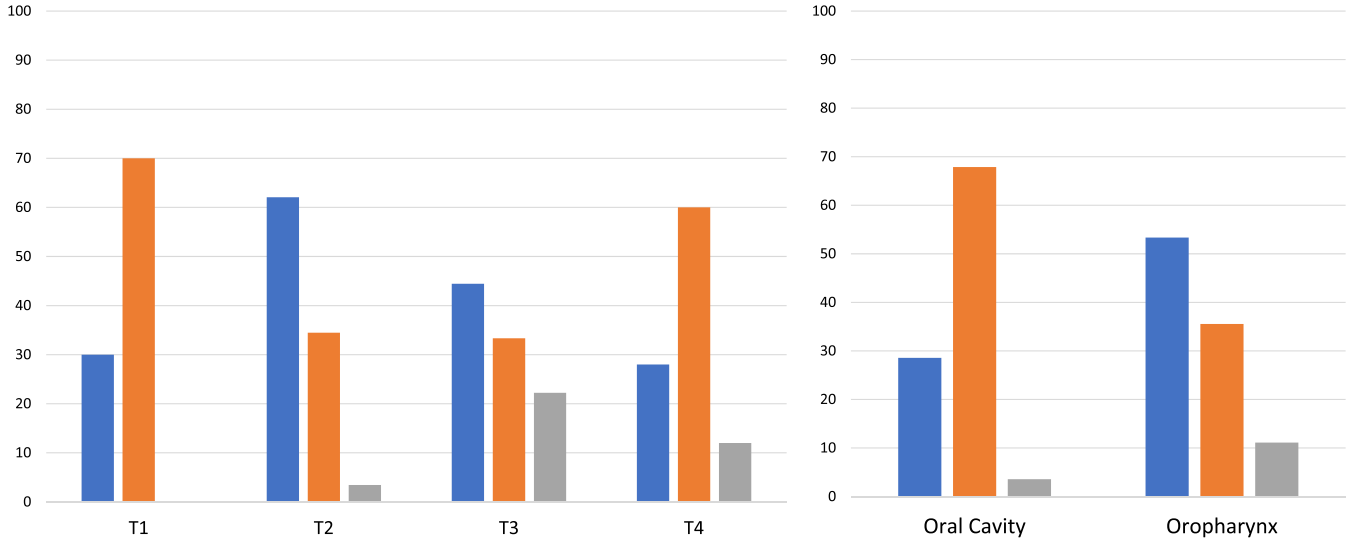

FIGURE 6 Severity profile (blue: light to mild [12–7], orange: moderate [<7–3], and severe: gray [>3–0]) of speech production disorders in our patient sample expressed in percentage: at the right by the size of the tumor, at the left by the localization between the oral cavity and oropharynx [Color figure can be viewed at wileyonlinelibrary.com]

communication field, which suggests a severity threshold of 3 with significant consequences on social life.

For illustrative purposes therefore, we can propose the following scale:

- between the maximum and 7, mild disorder at the limits of normal,
- between 7 and 3, moderate disorder with a low impact on comprehensibility and/or QoL relative to speech,
- inferior or equal to 3, severe disorder involving a loss of capacity for oral communication.

The representation of severity by creating levels of severity from the above determined thresholds is shown in Figure 6 according to the size and location of the tumor. The importance of tumor size is well illustrated. While there are more disorders present in cancers of the oral cavity, the most severe effects concern locations in the oropharynx in our population sample.

4 | DISCUSSION

In our population of patients treated for cancer of the UADT, the automatic scores measuring speech disorder show good psychometric properties, with parameters describing the severity of speech production disorders, while being much more reproducible than those obtained

from a human jury owing to automation of the analysis. This result is comparable to that obtained by other teams in oncology or neurology,[12,15–18,48–51] although those studies did not lead to the production of tools widely distributed in current clinical practice. This situation persists despite publications and developments dating back more than 10 years by certain teams, e.g. PEAKS developed on a continuous basis by the team of Erlangen in Germany.[7,16] One of the reasons may be the constant evolution of informatics and the explosive accelerated development of machine learning approaches in the automatic speech processing field. Thus, the work that we began in 2014 has already evolved in our team toward reprocessing of the data with deep neural networks. But before discussing this point, we would like to insist on two other aspects related to our work that may explain the limited use of results obtained up to now.

The first is the lack of integration in the research of complementary data aimed at conferring a clinical sense on the measurement of speech deficit. Indeed, by analogy with the measurement of auditory deficit, the hearing loss is interpreted in terms of the probability of the loss of comprehension of speech by the listener.[19] However, this approach is still missing for disorders of speech production. The prominent place of QoL questionnaires related to cancer to describe functional results, do not allow a solution to this problem. In fact, the concept of QoL, including that related to the state of health (HR-Qol) reflects the unique personal perception of health, taking into account the social dimension, functional and psychological factors. The functional factors are combined in the main questionnaires used in oncology of the UADT.[52] That is why the different components of the models of QoL must be explored between the deficits or the lesions and the QoL. Following the model of Wilson and Cleary[53] revisited by Ferrans et al.,[54] we explored the interactions between the different linguistic levels contributing to oral communication by using a measure of comprehensibility and a questionnaire resembling a functional status questionnaire (functional health status) which itself addresses the ability to perform tasks in connection with function and their effects on daily life. This work demonstrates the possibility to measure automatically and to classify the severity of speech disorders after the treatment in three categories, that is, mild, moderate, and severe.

The second point is the variability of perceptual evaluations used in reference to modeling of intelligibility. The review of the different publications cited above brings out several pitfalls that the C2SI has tried to avoid. The first is the need to use listener juries with the "burden" that that represents in terms of number of listeners. In fact, although there is no formal rule, a minimum of five listeners is justified to increase the reliability of the measurement, given the high interjudge variability.[55] The second is the type of task required, but also the instructions given for the same type of task, which are decisive.[19] This explains that few teams have been able to evaluate on a same patient population, different levels of integration of speech production in the process of oral communication. The most frequently used reference is the intelligibility measurement, either by a visual analog word identification scale or by a Likert scale[56] These tasks are indeed well correlated between each other, but differences may arise from a lack of agreement between the judges. This information is often missing and even in well conducted studies, agreements are moderate (0.55 in the study of Windrich et al.,[15] 0.69 in our study on the same type of task). Imperfection of the reference is, therefore, a limit to consider in the expected results and it confirms the importance of using automatic systems.

These pitfalls complicate the application of methods enabling interpretation of the scores obtained. Thus, the ROC curves used to validate the diagnostic tests are limited in this application by the uncertainty of the normality thresholds. Our results still need to be validated by continuing work on the tests of comprehension and of oral communication evaluation in the population of patients with UADT cancers, and on the methods for integrating these data in the validation of tools for measurement of speech deficits. That is why our first results are given for information purposes. In addition, the number of patients in our population is insufficient to carry out a subgroup analysis integrating the main factors influencing the results, such as size and location of the tumor, treatment modalities. This limitation does not allow to explore the impact of treatment modalities according to tumor location and particularly the impact of surgical treatments requiring reconstruction.

The different points that we have raised remain relevant as the development of artificial intelligence is in the process of improving results already obtained for the modeling of intelligibility. Different types of neural networks and neural network architectures are being evaluated. That especially concerns speech production disorders in neurology, with results already published that, while still insufficient, are very promising,[57–60] but also in UADT oncology.[61] It is in this perspective that the C2SI corpus[20] is used in the RUGBI project*: Looking for Relevant linguistic Units to improve the intelliGiBIlity measurement of speech production disorders.

In any case, the development of a device on a tablet allowing use in clinical practice is being developed. It will allow us to evaluate the C2SI score and the averaged normalized likelihood scores on another sample of patients treated for a cancer of the upper aerodigestive tract and to pursue its validations.

To close this article, this index could be applied to any language where linguistic resources (speech database, electronic dictionary, acoustic, and language models) are available for automatic speech processing tools. It will require to use a reference text in the target language), to construct a pseudoword repetition task respecting the phonotactic rules of the target language and then to study its performance according to the methodology described above.

5 | CONCLUSION

Our work made it possible to evaluate the validity of an automatic score of the measurement of speech disorders in patients treated for a cancer of the upper aerodigestive tract. The automatization of analyses allows a perfect reproducibility of scores. These scores may be considered for use in clinical practice because they are based on tasks already performed daily by therapists and are based on easily accessible automatic tools. The most contributive parameters are the automatic averaged normalized likelihood score on a reading text and the score of cumulative ranks on pseudowords. The correlation is 0.87 with the perceptual reference score, 0.77 with the comprehension score, and 0.5 with the QoL related to speech.

ACKNOWLEDGMENTS

This work has been carried out thanks to the French National Cancer Institute (INCa: Institut National du Cancer) in the 2014 C2SI project (Grant No.INCa-SHS-2014-135), and to the French National Research Agency (ANR: Agence national de la Recherche) in the 2018 untitled RUGBI project "Improving the measurement of intelligibility of pathological production disorders impaired speech" (Grant No. ANR-18-CE45-0008).

ENDNOTE

* https://www.irit.fr/rugbi/.

DATA AVAILABILITY STATEMENT

Data available on request due to privacy/ethical restrictions.

ORCID

Virginie Woisard https://orcid.org/0000-0003-3895-2827
Mathieu Balaguer https://orcid.org/0000-0003-1311-4501
Jérôme Farinas https://orcid.org/0000-0002-7456-9019
Alain Ghio https://orcid.org/0000-0001-7302-0799
Muriel Lalain https://orcid.org/0000-0002-7672-8589
Corine Astesano https://orcid.org/0000-0002-0882-4974
Julien Pinquier https://orcid.org/0000-0003-1556-1284
Benoît Lepage https://orcid.org/0000-0002-1586-2055

REFERENCES

1. Mlynarek AM, Rieger JM, Harris JR, et al. Methods of functional outcomes assessment following treatment of oral and oropharyngeal cancer: review of the literature. *J Otolaryngol Head Neck Surg*. 2008;37:2-10.
2. Borggreven PA, Verdonck-de Leeuw IM, Muller MJ, et al. Quality of life and functional status in patients with cancer of the oral cavity and oropharynx: pretreatment values of a prospective study. *Eur Arch Otorhinolaryngol*. 2007;264:651-657.
3. Stelzle F, Knipfer C, Schuster M, et al. Factors influencing relative speech intelligibility in patients with oral squamous cell carcinoma: a prospective study using automatic, computer-based speech analysis. *Int J Oral Maxillofac Surg*. 2013;42:1377-1384.
4. Colangelo LA, Logemann JA, Rademaker AW. Tumor size and pretreatment speech and swallowing in patients with resectable tumors. *Otolaryngol Head Neck Surg*. 2000;122:653-661.
5. Barrett WL, Gluckman JL, Wilson KM, Gleich LL. A comparison of treatments of squamous cell carcinoma of the base of tongue: surgical resection combined with external radiation therapy, external radiation therapy alone, and external radiation therapy combined with interstitial radiation. *Brachytherapy*. 2004;3:240-245.
6. Matsui Y, Ohno K, Yamashita Y, Takahashi K. Factors influencing postoperative speech function of tongue cancer patients following reconstruction with fasciocutaneous/myocutaneous flaps—a multicenter study. *Int J Oral Maxillofac Surg*. 2007;36:601-609.
7. Stelzle F, Oetter N, Goellner LT, et al. Speech intelligibility in patients with oral cancer: an objective baseline evaluation of pretreatment function and impairment. *Head Neck*. 2019;41:1063-1069.
8. Dwivedi RC, St Rose S, Roe JW, et al. First report on the reliability and validity of speech handicap index in native English-speaking patients with head and neck cancer. *Head Neck*. 2011;33:341-348.
9. Gupta B, Johnson NW, Kumar N. Global epidemiology of head and neck cancers: a continuing challenge. *Oncology*. 2016;91:13-23.
10. Balaguer M, Pommée T, Farinas J, Pinquier J, Woisard V, Speyer R. Effects of oral and oropharyngeal cancer on speech intelligibility using acoustic analysis: systematic review. *Head Neck*. 2020;42:111-130.
11. Van Nuffelen G, De Bodt M, Vanderwegen J, Van de Heyning P, Wuyts F. Effect of rate control on speech production and intelligibility in dysarthria. *Folia Phoniatr Logop*. 2010;62:110-119.
12. Hustad KC, Oakes A, Allison K. Variability and diagnostic accuracy of speech intelligibility scores in children. *J Speech Lang Hear Res*. 2015;58:1695-1707.
13. Fex S. Perceptual evaluation. *J Voice*. 1992;6:155-158.
14. Berisha V, Utianski R, Liss J. Towards a clinical tool for automatic intelligibility assessment. *Proc IEEE Int Conf Acoust Speech Signal Process*. 2013;2825-2828.

15. Windrich M, Maier A, Kohler R, Nöth E, et al. Automatic quantification of speech intelligibility of adults with oral squamous cell carcinoma. *Folia Phoniatr Logop*. 2008;60:151-156.
16. Maier A, Haderlein T, Eysholdt U, et al. PEAKS—A system for the automatic evaluation of voice and speech. *Speech Commun*. 2009;55:425-437.
17. Middag C, Clapham R, Van Son R, Martens JP. Robust automatic intelligibility assessment techniques evaluated on speakers treated for head and neck cancer. *Comput Speech Lang*. 2014;28:467-482.
18. Van Nuffelen G, Middag C, De Bodt M, Martens JP. Speech technology-based assessment of phoneme intelligibility in dysarthria. *Int J Lang Commun Disord*. 2009;44:716-730.
19. Woisard V, Lepage B. Difference between the degree of intelligibility and the degree of severity perception of speech disorders. *Audiol Med*. 2010;00:1-8.
20. Woisard V, Astésano C, Balaguer M, et al. C2SI corpus: a database of speech disorder productions to assess intelligibility and quality of life in head and neck cancers. *Lang Resour Eval*. 2020;55:173-190. doi:10.1007/s10579-020-09496-3
21. Lalain M, Ghio A, Giusti L, Robert D, Fredouille C, Woisard V. Design and development of a speech intelligibility test based on pseudowords in French: why and how? *J Speech Lang Hear Res*. 2020;63:1-14.
22. Ghio A, Pouchoulin G, Teston B, et al. How to manage sound, physiological and clinical data of 2500 dysphonic and dysarthric speakers? *Speech Commun*. 2012;54:664-679.
23. Balaguer M, Boisguérin A, Galtier A, Gaillard N, Puech M, Woisard V. Assessment of impairment of intelligibility and of speech signal after oral cavity and oropharynx cancer. *Eur Ann Otorhinolaryngol Head Neck Dis*. 2019;136:355-359.
24. Dittner J, Lepage B, Woisard V, et al. Construction and validation of a quantitative assessment of speech intelligibility for speech disorders. *Rev Laryngol Otol Rhinol*. 2010;131:9-14.
25. Pisoni DB, Manous LM, Dedina MJ. Comprehension of natural and synthetic speech: effects of predictability on the verification of sentences controlled for intelligibility. *Comput Speech Lang*. 1987;2:303-320.
26. Nocaudie O, Astésano C, Ghio A, Lalain M & Woisard V Evaluation de la compréhensibilité et conservation des fonctions prosodiques en perception de la parole de patients post-traitement de cancers de la cavité buccale et du pharynx. XXXIIe Journées Etudes Sur la Parole; 2018:196-120.
27. Brierley JD, Gospodarowicz MK, Wittekind C, *TNM Classification of Malignant Tumours*. Vol 1. 8. Wiley Blackwell; 2017.
28. Rinkel RN, Verdonck-de Leeuw IM, van Reij EJ, Aaronson NK, Leemans CR. Speech handicap index in patients with oral and pharyngeal cancer: better understanding of patients' complaints. *Head Neck*. 2008;30:868-874.
29. Fichaux-Bourin P, Woisard V, Grand S, Puech M, Bodin S. Validation of a self assessment for speech disorders (phonation handicap index). *Rev Laryngol Otol Rhinol*. 2009;130:45-51.
30. Balaguer M, Farinas J, Fichaux-Bourin P, Puech M, Pinquier J, Woisard V. Validation of the French versions of the speech handicap index and the phonation handicap index in patients treated for cancer of the oral cavity or oropharynx. *Folia Phoniatr Logop*. 2019;72(6):464-477. doi:10.1159/000503448
31. Kent RD, Weismer G, Kent JF, Rosenbek JC. Toward phonetic intelligibility testing in dysarthria. *J Speech Hear Disord*. 1989;54:482-499.
32. Keintz CK, Bunton K, Hoit JD. Influence of visual information on the intelligibility of dysarthric speech. *Am J Speech Lang Pathol*. 2007;16:222-234.
33. Lindblom B. On the communication process: speaker-listener interaction and the development of speech. *Augment Altern Commun*. 1990;6:220-230.
34. Yorkston KM, Strand EA, Kennedy MRT. Comprehensibility of dysarthric speech: implications for assessment and treatment planning. *Am J Speech Lang Pathol*. 1996;5:55-65.
35. Auzou P. Les objectifs du bilan de la dysarthrie. *Les Dysarthries*. Solal; 2007:189-195.
36. Balaguer M, Boisguerin A, Galtier A, Gaillard N, Puech M, Woisard V. Factors influencing intelligibility and severity of chronic speech disorders of patients treated for oral or oropharyngeal cancer. *Eur Arch Otorhinolaryngol*. 2019;276:1767-1774.
37. Ohala J. Phonological evidence for top-down processing in speech perception. In: Erlbaum L. *Invariance and Variability in Speech Processes*. Perkell & Klatt; 1986. p 386–397.
38. Eyben F. *Real-Time Speech and Music Classification by Large Audio Feature Space Extraction*. Springer; 2016:298.
39. Paul B. Praat, a system for doing phonetics by computer. *Glot Int*. 2001;5:341-345.
40. Menin Sicard A, Sicard E. *Evaluation and Rehabilitation of the Voice—Clinical and Objective Approach*. De Boeck Supérieur; 2016:233.
41. Laaridh I, Fredouille C, Ghio A, Lalain M, Woisard V. Automatic Evaluation of Speech Intelligibility Based on i-Vectors in the Context of Head and Neck Cancers. Proceedings of Interspeech'18, Hyderabad, India; 2018.
42. Laaridh I, Meunier C, Fredouille C. Perceptual evaluation for automatic anomaly detection in disordered speech: focus on ambiguous cases. *Speech Commun*. 2018;105:23-33.
43. Fredouille C, Ghio A, Laaridh I, Lalain M & Woisard V Acoustic-Phonetic Decoding for Speech Intelligibility Evaluation in the Context of Head and Neck Cancers. International Congress of Phonetic Sciences (ICPhS), Melbourne, Australia; 2019.
44. Ghio A, Lalain M, Giusti L, Fredouille C, Woisard V. How to Compare Automatically Two Phonological Strings: Application to Intelligibility Measurement in the Case of Atypical Speech. 12th Conference on Language Resources and Evaluation (LREC 2020), Marseille, France: ELRA; 2020.
45. StataCorp. *Stata Statistical Software: Release 14*. StataCorp LP; 2015.
46. Mukaka MM. Statistics corner: a guide to appropriate use of correlation coefficient in medical research. *Malawi Med J*. 2012;24:69-71.
47. Bouter LM, Zielhuis GA, Zeegers MPA. Diagnostic and prognostic research. *Textbook of Epidemiology*. Bohn Stafleu Van Loghum; 2007:171-201.
48. Khan T, Westin J, Dougherty M. Classification of speech intelligibility in Parkinson's disease. *Biocybern Biomed Eng*. 2014;34:35-45.
49. Laaridh I, Fredouille C, Meunier C. Automatic detection of phone-based anomalies in dysarthric speech. *ACM Trans Access Comput*. 2015;6(3):1-24. doi:10.1145/2739050

50. Kim J, Kumar N, Tsiartas A, Li M, Narayanan SS. Automatic intelligibility classification of sentence-level pathological speech. *Comput Speech Lang*. 2015;29:132-144.
51. Vásquez-Correa JC, Orozco-Arroyave JR, Bocklet T, Nöth E. Towards an automatic evaluation of the dysarthria level of patients with Parkinson's disease. *J Commun Disord*. 2018;76:21-36.
52. Murphy BA, Ridner S, Wells N, Dietrich M. Quality of life research in head and neck cancer: a review of the current state of the science. *Crit Rev Oncol Hematol*. 2007;62:251-267.
53. Wilson IB, Cleary PD. Linking clinical variables with health-related quality of life: a conceptual model of patient outcomes. *JAMA*. 1995;273:59-65.
54. Ferrans CE, Zerwic JJ, Wilbur JE, Larson JL. Conceptual model of health-related quality of life. *J Nurs Scholarsh*. 2005; 37:336-342.
55. Norme ISO TR 4870. Acoustique—Élaboration et étalonnage des tests d'intelligibilité de parole; 1991.
56. Hustad KC. The relationship between listener comprehension and intelligibility scores for speakers with dysarthria. *J Speech Lang Hear Res*. 2008;51:562-573.
57. Kim MJ, Cao B, An K & Wang J Dysarthric Speech Recognition Using Convolutional LSTM Neural Network. Proceedings of Interspeech'18, Hyderabad, India; 2018.
58. Wang J, Kothalkar PV, Kim M, et al. Automatic prediction of intelligible speaking rate for individuals with ALS from speech acoustic and articulatory samples. *Int J Speech Lang Pathol*. 2018;20:669-679.
59. Vasquez-Correa JC, Arias-Vergara T, Orozco-Arroyave JR, Eskofier B, Klucken J, Noth E. Multimodal assessment of Parkinson's disease: a deep learning approach. *IEEE J Biomed Health Inform*. 2019;23:1618-1630.
60. Chandrashekar HM, Karjigi V, Sreedevi N. Spectro-temporal representation of speech for intelligibility assessment of dysarthria. *IEEE J Sel Top Signal Process*. 2020;14:390-399.
61. Bin L, Kelley MC, Aalto D & Tucker BV Automatic Speech Intelligibility Scoring of Head and Neck Cancer Patients with Deep Neural Networks. International Congress of Phonetic Sciences ICPhS; 2019. https://icphs2019.org/icphs2019-fullpapers/pdf/full-paper_448.pdf.

How to cite this article: Woisard V, Balaguer M, Fredouille C, et al. Construction of an automatic score for the evaluation of speech disorders among patients treated for a cancer of the oral cavity or the oropharynx: The Carcinologic Speech Severity Index. *Head & Neck*. 2021;1-18. doi:10.1002/hed.26903