Contents lists available at ScienceDirect

European Journal of Radiology

Review

Intrarater and interrater reliability for measurements in videofluoroscopy of swallowing

Laura Baijens [a,*], Ali Barikroo [b], Walmari Pilz [a]

[a] Department of Otorhinolaryngology, Head and Neck Surgery, Maastricht University Medical Center, Maastricht, The Netherlands
[b] Swallowing Research Laboratory, Department of Speech, Language and Hearing Sciences, College of Public Health and Health Professions, University of Florida, Gainesville, FL, USA

ARTICLE INFO

ABSTRACT

Objective: Intrarater and interrater reliability is crucial to the quality of diagnostic and therapy-effect studies. This paper reports on a systematic review of studies on intrarater and interrater reliability for measurements in videofluoroscopy of swallowing. The aim of this review was to summarize and qualitatively analyze published studies on that topic.
Materials and methods: Those published up to March 2013 were found through a comprehensive electronic database search using PubMed, Embase, and The Cochrane Library. Two reviewers independently assessed the studies using strict inclusion criteria.
Results: Nineteen studies were included and then qualitatively analyzed. In several of these, methodological problems were found. Moreover, intrarater and interrater reliability varied with the measure applied. A meta-analysis was not carried out as studies were not of sufficient quality to warrant doing so.
Conclusion: In order to achieve reliable measurements in videofluoroscopy of swallowing, it is recommended that raters use well-defined guidelines for the levels of ordinal visuoperceptual variables. Furthermore, in order to make the measurements reliable (intrarater and interrater) it is recommended that, following protocolled pre-experimental training, the raters should have maximum consensus about the definition of the measured variables.

© 2013 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

The measurements used to evaluate swallowing function in videofluoroscopy of swallowing (VFS) are not always reliable, which gives cause for concern. If intrarater and interrater reliability is not good, the VFS measurements cannot be trusted as grounds to evaluate either swallowing function or the outcome of dysphagia treatment. In each new study, the measurements of swallowing variables – for instance, ordinal visuoperceptual, continuous temporal, or continuous spatial variables – must be reliable enough to warrant statistical analysis. Any study of swallowing should provide information on the training of the raters and the protocol to assess intrarater and interrater reliability. Only then can the study outcome be accurately interpreted. The present paper offers a summary and qualitative analysis of the published studies on intrarater and interrater reliability for measurements in VFS. The research question for this systematic review was: How could measurements in future VFS studies be made more reliable?

2. Methods

2.1. Identification and selection of studies

Two of the present authors independently carried out the literature search until March 2013 using the electronic databases Embase, PubMed, and The Cochrane Library. These were chosen because they are the most relevant biomedical databases for clinicians involved in swallowing assessment. The specific limits were publications in the Dutch, English, French, German, or Spanish language. In Embase, Thesaurus terms were combined as presented in Table 1. All trees of subject headings were explored. The search was expanded by using free-text words (truncation or wildcards). In PubMed and in The Cochrane Library database, Medical Subject Heading (MeSH) terms were recognized and all trees of these MeSH terms were explored and combined in the biomedical literature engines (Table 1). The free-text words used in Embase were explored in PubMed and in The Cochrane Library database (Table 1). The reference lists of all included articles were searched for additional studies. This search did not yield any additional articles.

* Corresponding author at: Department of Otorhinolaryngology, Head and Neck Surgery, Maastricht University Medical Center, P.O. Box 5800, 6202 AZ Maastricht, The Netherlands. Tel.: +31 433875585; fax: +31 433875580.
  E-mail addresses: laura.baijens@mumc.nl (L. Baijens), a.Barikroo@ufl.edu (A. Barikroo), walmari.pilz@mumc.nl (W. Pilz).

Table 1
The systematic syntax of the extended literature search in the three biomedical engines.

Embase: Thesaurus terms
[*reproducibility* OR *validation process* OR *reliability* OR *observer variation*] AND [*swallowing* OR *dysphagia* OR *fluoroscopy* OR *barium*]

Embase: Free-text words
[*swallow** OR *dysphag** OR *deglutit**] were combined with [*reliab** OR *valid** OR *variat** OR *variabil** OR *reproducib** OR *intrarat** OR *interrat** OR *intraob** OR *interob** OR *intrajud** OR *interjud** OR *agree**] AND [*fluoroscop** OR *video** OR *barium**]

PubMed: Medical Subject Heading (MeSH) terms
[*Reproducibility of Results* OR *Validation Studies as Topic* OR *Validation Studies* OR *Observer Variation*] AND [*Deglutition* OR *Deglutition Disorders* OR *Fluoroscopy* OR *Barium* OR *Barium Sulfate*]

PubMed: Free-text words
[*swallow** OR *dysphag** OR *deglutit**] were combined with [*reliab** OR *valid** OR *variat** OR *variabil** OR *reproducib** OR *intrarat** OR *interrat** OR *intraob** OR *interob** OR *intrajud** OR *interjud** OR *agree**] AND [*fluoroscop** OR *video** OR *barium**]

The Cochrane Library database: Medical Subject Heading (MeSH) terms
[*reproducibility of results* OR *validation studies as topic* OR *observer variation*] AND [*deglutition* OR *deglutition disorders* OR *fluoroscopy* OR *barium*]

The Cochrane Library database: Free-text words
[*swallow** OR *dysphag** OR *deglutit**] were combined with [*reliab** OR *valid** OR *variat** OR *variabil** OR *reproducib** OR *intrarat** OR *interrat** OR *intraob** OR *interob** OR *intrajud** OR *interjud** OR *agree**] AND [*fluoroscop** OR *video** OR *barium**]

Studies were included that describe experiments to specifically evaluate and/or improve intrarater and interrater reliability for measurements in VFS examinations (Table 2). Studies describing several measurement tools were included if they covered the reliability analyses for the VFS measurements [1,2]. Abstracts that were anecdotal, speculative, or editorial in nature were not included. Two reviewers independently based their first selection of studies on the abstracts. The very few differences in their search were resolved using consensus agreement following discussion. Also excluded was the gray literature for the reason that basic information such as authorship, publication date, or publishing body may not be discerned with certainty.

2.2. Data analysis and assessment of study quality

No validated instrument is available for assessing the methodological quality of studies on intrarater and interrater reliability. Therefore, a list of criteria for quality assessment was compiled, as derived from the studies of Katrak et al., Reitsma et al., and Whiting et al. [3–5]. Data extraction consisted of the analysis of critical appraisal items per included study. The items evaluated are pre-

Table 2
Inclusion criteria for studies.

Inclusion criteria
Design
Repeated measurements within and/or between raters
Peer-reviewed journal articles
Dutch, English, French, German, or Spanish language articles

Dysphagic participants
Symptomatic and asymptomatic subjects
Adults

Measurements in videofluoroscopy
Ordinal visuoperceptual variables
Continuous temporal variables
Continuous spatial variables

Outcome
Intrarater and/or interrater reliability

Table 3
Critical appraisal criteria for methodological quality assessment [3–5].

---

1. Sample selection. Was a representative sample of participants used?
2. Was a representative sample of raters used?
3. Is replication of the measurement procedure possible?
4. Was clinical patient information unavailable to the raters?
5. Was randomization of subjects' swallows performed?
6. Were study aims and justification reported?
7. Were subjects' characteristics stable during the study?
8. Were raters' characteristics stable during the study?
9. Patient attrition. Was an explanation for drop-outs provided?
10. Were the raters blinded to each other's results?
11. Was an estimate of intrarater reliability determined?
12. Was an estimate of interrater reliability determined?
13. Method of data analyses. Were appropriate measures (kappa, ICC) used for calculating reliability?

---

sented in Table 3. They were rated as 'yes', 'no', or as 'unknown' when insufficient information was provided (Table 4). Criteria 1–4 were used to assess external validity, criteria 5–12 to assess internal validity (determining risk of bias), and criterion 13 was used to assess statistical methods [6]. The main criterion of external validity is the process of generalization, specifically whether results obtained from a small sample can be extended to make predictions about the entire population. Internal validity is an inductive estimate of the degree to which conclusions can be made (e.g., cause and effect) based on the measures used and the whole research design [7]. The present quality assessment tool, like many other validated ones, does not incorporate a quality score [5,8]. Choices on how to weight and calculate quality scores are generally fairly arbitrary, so it would be impossible to generate an objective quality score [5]. Two reviewers rated the critical appraisal items independently [6]. They were not blinded to the journal and the authors. The very few differences in their rating were settled by discussion and consensus agreement. A meta-analysis was not carried out as studies were not of sufficient quality to warrant doing so.

3. Results

3.1. General results

In total, 3952 articles were found. Using Thesaurus terms, 1243 articles were selected in Embase. Using MeSH terms, 1466 articles were selected in PubMed and 68 in The Cochrane Library. The search using free-text words turned up 659 articles in Embase, 495 articles in PubMed, and 21 in The Cochrane Library. A total of 35 articles were selected for full-text review. Overlap between the biomedical databases and overlap between free-text, MeSH, and Thesaurus terms was excluded, leaving 19 articles for inclusion and subsequent review (Fig. 1).

3.2. Brief description of studies on intrarater and interrater reliability

Table 5 summarizes the data per study (if present). Each one was reviewed for the following methodological items as well: etiology of dysphagia; number of swallows per subject; swallow protocol during VFS examination; number of repeated measurements for intrarater reliability; number of compared measurements for interrater reliability; pre-experimental training program for raters; the rating task; speed of video for analysis (frame-by-frame, real-time speed, etc.); and randomization of the analyzed swallows. The studies are listed in alphabetical order of the first author's name. The first 14 articles studied reliability for measurements in VFS as their primary goal. The final five articles studied observer reliability as a secondary goal.