diagnostics

MDPI

Article

Radiological Reporting of Brain Atrophy in MRI: Real-Life Comparison Between Narrative Reports, Semiquantitative Scales and Automated Software-Based Volumetry

Federico Bruno [1,2,*], Cristina Fagotti [1], Gaspare Saltarelli [1], Giovanni Di Cerbo [1], Alessandra Sabatelli [1], Claudia De Felici [1], Antonio Innocenzi [1], Ernesto Di Cesare [1] and Alessandra Splendiani [1]

[1] Department of Biotechnological and Applied Clinical Sciences, University of L'Aquila, 67100 L'Aquila, Italy
[2] Neuroradiology, San Salvatore Hospital, 67100 L'Aquila, Italy
* Correspondence: federico.bruno@univaq.it

Abstract: Background: Accurate assessment of brain atrophy is essential in the diagnosis and monitoring of brain aging and neurodegenerative disorders. Radiological methods range from narrative reporting to semi-quantitative visual rating scales (VRSs) and fully automated volumetric software. However, their integration and consistency in clinical practice remain limited. **Methods**: In this retrospective study, brain MRI images of 43 patients were evaluated. Brain atrophy was assessed by extrapolating findings from narrative radiology reports, three validated VRSs (MTA, Koedam, Pasquier), and Pixyl.Neuro.BV, a commercially available volumetric software platform. Agreement between methods was assessed using intraclass correlation coefficients (ICCs), Cohen's kappa, Spearman's correlation, and McNemar tests. **Results**: Moderate correlation was found between narrative reports and VRSs (ρ = 0.55–0.69), but categorical agreement was limited (kappa = 0.21–0.30). Visual scales underestimated atrophy relative to software (mean scores: VRSs = 0.196; software = 0.279), while reports tended to overestimate. Agreement between VRSs and software was poor (kappa = 0.14–0.33), though MTA showed a significant correlation with hippocampal volume. Agreement between reports and software was lowest for global atrophy. **Conclusions**: Narrative reports, while common in practice, show low consistency with structured scales and quantitative software, especially in subtle cases. VRSs improve standardization but remain subjective and less sensitive. Integrating structured scales and volumetric tools into clinical workflows may enhance diagnostic accuracy and consistency in dementia imaging.

Keywords: MRI; dementia; brain atrophy; brain volumetry; automated software

Academic Editor: Andreas Kjaer

Received: 2 April 2025
Revised: 3 May 2025
Accepted: 6 May 2025
Published: 14 May 2025

Citation: Bruno, F.; Fagotti, C.; Saltarelli, G.; Di Cerbo, G.; Sabatelli, A.; De Felici, C.; Innocenzi, A.; Di Cesare, E.; Splendiani, A. Radiological Reporting of Brain Atrophy in MRI: Real-Life Comparison Between Narrative Reports, Semiquantitative Scales and Automated Software-Based Volumetry. *Diagnostics* **2025**, *15*, 1246. https://doi.org/10.3390/diagnostics15101246

1. Introduction

Brain atrophy is a hallmark of many neurodegenerative diseases and plays a pivotal role in understanding their underlying pathophysiology [1–6]. Magnetic resonance imaging (MRI) has become an essential tool for detecting and monitoring brain atrophy, providing detailed visualization of affected regions and enabling quantification of tissue loss. Structural MRI, in particular, is widely used to evaluate atrophy patterns associated with various forms of dementia, aiding both in diagnosis and longitudinal disease tracking. Imaging biomarkers—such as regional volume loss—are crucial for improving early detection and differential diagnosis of neurodegenerative conditions. Volumetric MRI, multiparametric approaches, and molecular imaging techniques have further enhanced diagnostic sensitivity and specificity [2,7–12].

MRI has become increasingly important in characterizing primary neurodegenerative pathology, not only by detecting regional brain atrophy but also by identifying other structural markers associated with aging and neurodegeneration. These include white matter hyperintensities (WMHs), cerebral microbleeds, and enlarged perivascular spaces, which provide valuable insights into underlying small vessel disease and overall brain health. The detection of these features enhances the diagnostic accuracy for various dementia subtypes and contributes to a more comprehensive assessment of brain aging [8,13–18].

Importantly, brain atrophy is not exclusive to pathology: it is also a feature of normal aging, which is typically associated with diffuse and gradual brain volume reduction. However, unlike the selective and region-specific patterns of neurodegeneration, age-related atrophy tends to occur more uniformly and progresses at a slower rate. Distinguishing between physiological tissue loss and pathological atrophy is particularly critical in the context of today's aging population, where increasing life expectancy is contributing to a higher prevalence of cognitive decline and dementia-related conditions. Accurate differentiation is essential for timely diagnosis, appropriate clinical management, and avoiding both over- and underdiagnosis in older individuals [17,19–22].

To improve interpretability and consistency, semi-quantitative visual rating scales (VRSs) such as the Medial Temporal Atrophy (MTA) scale, Global Cortical Atrophy (GCA) Scale, and the Fazekas scale for white matter hyperintensities have been introduced. These tools enable structured evaluation of imaging findings, but their use in routine clinical practice remains variable [8,13,15,16].

The lack of standardization in radiological reporting presents a significant challenge in both clinical practice and research. Unstructured or free-text reports are still widely used, despite their inherent limitations. These reports rely heavily on subjective interpretation and may omit essential details required for a comprehensive diagnostic assessment. In contrast, structured reporting has been proposed to enhance standardization. Structured reports utilize predefined templates tailored to specific diseases or clinical indications, ensuring systematic inclusion of all relevant information [23].

In the context of neurodegenerative disorders, variability in reporting is particularly problematic. Critical findings, such as the presence and extent of brain atrophy or white matter hyperintensities, may be under-reported or inconsistently described. Furthermore, the use of semi-quantitative scales is often inconsistent or absent. This lack of standardization can make it difficult to compare results between centers and to integrate imaging data into large-scale studies aimed at understanding disease progression or treatment response [9,24].

The concept of "contextual reporting" has emerged as a compromise between structured and free-text reporting, where the structure of the report is tailored to the specific pathology being evaluated, ensuring consistent capture of essential details while allowing the radiologist the freedom to provide individualized insights [25–27]. Radiological rating scales (VRSs) play a crucial role in this area, allowing for better recognition of radiological findings in diagnostic protocols for dementia. The literature demonstrates that the characteristic findings of brain atrophy are often underdiagnosed and not consistently reported in radiological reports. It is suggested that the personal experience of the radiologist and the predominant use of imaging to exclude secondary and reversible causes of dementia are key factors underlying the discrepancies observed between reports from different hospitals. Furthermore, the use of the VRS remains poorly documented. At the European level, 75% of centers utilize the VRS in clinical practice, 82% report changes in white matter, and only 6% regularly employ quantitative volumetric measurements [8,13,15,16]. Among the most widely used scales, the Medial Temporal Atrophy (MTA) scale is the most prevalent. The main barrier to widespread use of the VRS is the lack of specific training required to ensure a high level of intra-rater agreement [13].

In parallel, with the exploding diffusion of AI-based solutions in neuroradiology, the use of dedicated software for automated brain volumetry is becoming increasingly common in clinical practice [28–30]. These tools offer objective, reproducible measurements of brain structures and have the potential to complement visual assessments, particularly in early or subtle cases of atrophy [31–34].

This study aims to evaluate how brain atrophy is currently documented in clinical MRI reports and to assess the consistency of these reports with findings derived from visual rating scales and automated volumetric software. Specifically, we investigate the degree of concordance between narrative radiology reports, semi-quantitative visual scales, and software-based metrics, with the goal of understanding the potential complementary role of automated tools in the diagnostic workflow for neurodegenerative disease.

2. Materials and Methods

2.1. Study Population

This observational study was conducted on a cohort of 43 patients (10 women, 33 men, with a mean age of 67.65 $\pm$ 9.53 years) retrospectively selected from a cohort previously enrolled in a prior study conducted at our institution [35].

All patients had been scanned using a standardized MRI protocol to ensure consistency in image acquisition and comparability across the various methods of analysis.

Inclusion criteria consisted of the availability of a 3D T1-weighted volumetric sequence suitable for automated processing, high-quality structural MR images and an absence of significant motion artifacts.

Patients were excluded if automated segmentation failed or was deemed technically inappropriate due to artifacts, incomplete sequences, or abnormal anatomy that interfered with the software's algorithms.

2.2. MRI Acquisition Protocol

All MRI studies were performed on the same 3 Tesla scanner (MR750w, GE Healthcare, Chicago, IL, USA) using a 32-channel phased-array head coil. The standardized protocol included the following sequences:

Axial T2-weighted FLAIR: slice thickness 3.0 mm (gap 0.3 mm), TR 11,000 ms, frequency FoV 24 cm, phase FoV 0.8.

Axial T2 GRE*: slice thickness 3.0 mm (gap 0.3 mm), TR 960 ms, frequency FoV 26 cm, phase FoV 0.75.

Axial SWI: slice thickness 2.0 mm, frequency FoV 24 cm, phase FoV 0.85.

Axial DWI: slice thickness 3.0 mm (gap 0.3 mm), TR 10,550 ms, frequency FoV 26 cm, phase FoV 0.8.

Axial and coronal T2-weighted: slice thickness 3.0 mm (gap 0.3 mm), TR 7854 ms, frequency FoV 26 cm, phase FoV 0.8.

Volumetric T1-weighted 3D-IR-FSPGR: slice thickness 1 mm (no gap), TR 8.5 ms, frequency FoV 25.6 cm, phase FoV 0.8.

All image datasets were anonymized prior to analysis to ensure patient confidentiality in accordance with institutional and ethical guidelines.

2.3. Automated Brain Volumetry

Quantitative brain volumetry was conducted using Pixyl.Neuro, a CE-marked, commercially available software platform developed by Pixyl SA (Grenoble, France). This tool enables automated segmentation and analysis of brain volumes using high-resolution 3D T1-weighted images.

The software performs two main types of segmentation: tissue-level segmentation, which distinguishes gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF), and structure-level segmentation, which identifies and quantifies specific brain structures including the hippocampi, thalami, putamen, and cortical lobes.

Following segmentation, the software computes the absolute volumes of these structures (in milliliters) and compares them to an internal normative database adjusted for patient age and sex. This yields percentile rankings, which express each structure's volume relative to a healthy reference population.

An example of a Pixyl report is provided (Figure 1), illustrating the software's output, including global brain volumes, regional metrics, and percentile comparisons.

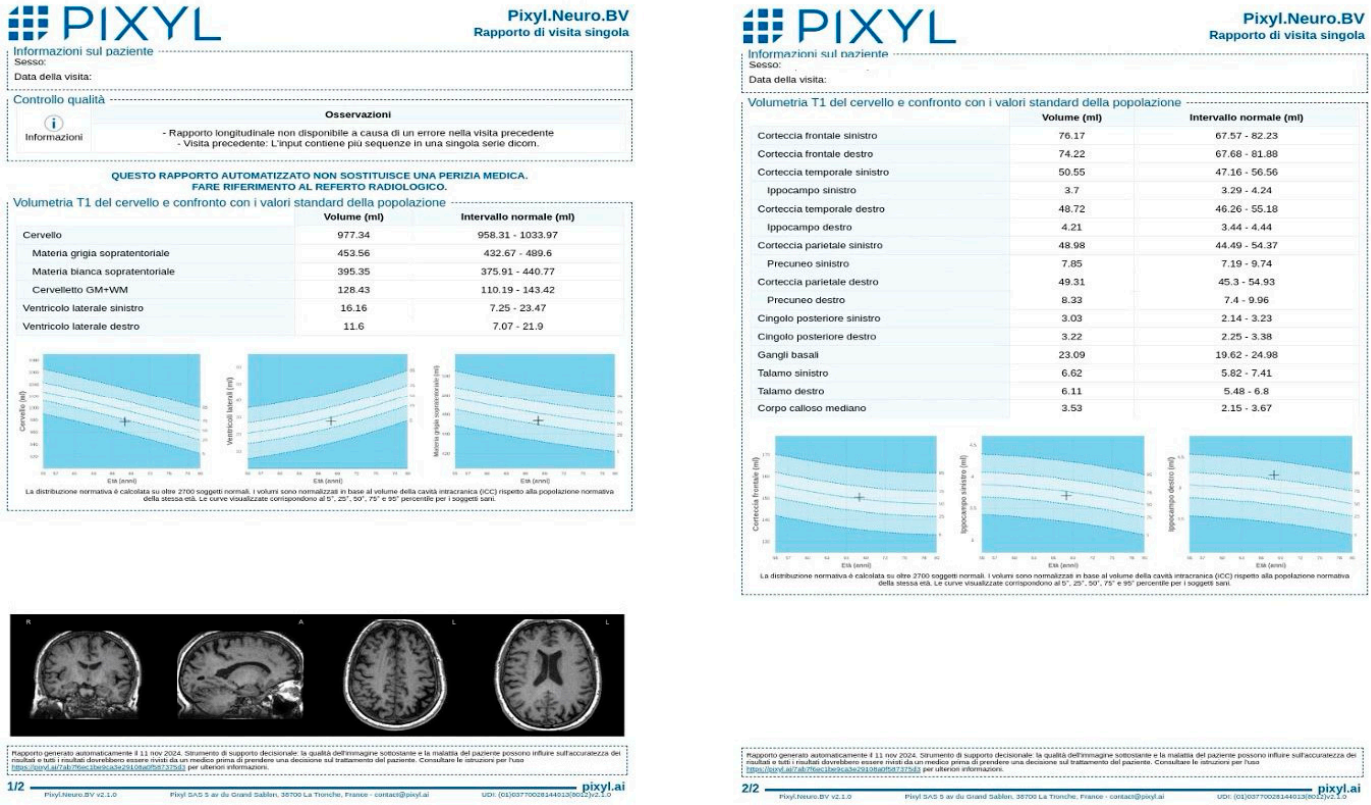

Figure 1. Example of Pixyl report after MRI brain volumetry, illustrating the software's quantitative volumetric measurements of brain structures. The report includes measures of global brain volume, regional volumes (e.g., hippocampus, ventricles), and comparisons to an age-matched normative database.

2.4. Visual Rating Scales (VRSs)

Each patient's MRI was independently evaluated by two trained neuroradiologists using three well-validated visual rating scales (VRSs) (Table 1).

Medial Temporal Atrophy (MTA) scale [36]: The MTA scale assesses atrophy in the medial temporal lobe structures, specifically focusing on the hippocampus, choroid fissure, and temporal horn of the lateral ventricle. Ratings range from 0 (no atrophy) to 4 (severe atrophy).

Scores ≥ 2 were considered abnormal for patients < 75 years; scores ≥ 3 were abnormal for patients ≥ 75 years.

Global Cortical Atrophy (GCA) Scale [37]: This scale evaluates sulcal widening and ventricular enlargement across all lobes. Each hemisphere is scored from 0 to 3, with higher scores indicating greater Global Cortical Atrophy. Scores of ≥ 2 were considered pathological in this study.

Table 1. A detailed summary of all three scales, including anatomical targets, imaging planes, scoring systems, and thresholds for pathological atrophy.

Scale	Region Assessed	Imaging Plane/Sequence	Scoring Range	Scoring Criteria	Pathological Cut-Off
MTA (Scheltens)	Medial temporal lobe (hippocampus, choroid fissure, temporal horn)	Coronal T1-weighted	0–4	0 = normal; 1 = mild choroid fissure widening; 2 = +mild temporal horn enlargement; 3 = +moderate hippocampal atrophy; 4 = severe atrophy with structural loss	≥2 (<75 yrs); ≥3 (≥75 yrs)
GCA (Pasquier)	Global Cortical Atrophy (frontal, parietal, temporal, occipital lobes)	Axial T1-weighted	0–3 (per hemisphere)	0 = normal; 1 = mild sulcal widening; 2 = moderate; 3 = severe "knife blade" atrophy	≥2 (any age)
Koedam	Posterior parietal regions (precuneus, parieto-occipital sulcus, posterior cingulate)	Axial, sagittal, coronal T1-weighted	0–3	0 = no sulcal widening; 1 = mild; 2 = moderate; 3 = severe widening and atrophy	≥2 (any age)

Posterior Atrophy scale of parietal atrophy [38]: Designed to assess parietal and posterior cingulate atrophy, this scale also uses a 0–3 range. A score of ≥2 was considered abnormal.

2.5. Qualitative Evaluation of Original Radiology Reports

Original narrative radiology reports corresponding to each MRI exam were reviewed independently to determine how the findings were described in routine clinical practice. Specifically, the reports were examined for any mention or description of medial temporal, global cortical, or posterior cortical atrophy—corresponding to the anatomical regions evaluated by the VRS.

A qualitative scoring system adapted from Torisson [4] was applied to classify how brain atrophy was documented in the original reports:

- NA: No mention of atrophy or related findings.
- 0: Atrophy explicitly reported as absent or normal.
- 1: Atrophy mentioned in vague or mild terms, without specific grading.
- 2: Atrophy described as moderate.
- 3: Atrophy described as severe.

This scoring allowed for a structured comparison between routine radiology practice (narrative reporting), visual rating by trained observers, and quantitative volumetric data generated by automated software.

2.6. Statistical Analysis

All statistical analyses were performed using MedCalc software version 23.1.6 (MedCalc Software Ltd., Ostend, Belgium). The primary objective of the analysis was to assess the degree of agreement and correlation between three different methods of brain atrophy assessment: narrative radiology reports, semi-quantitative visual rating scales, and automated volumetric analysis using Pixyl.Neuro.

Spearman's rho (ρ) was calculated to evaluate the correlation between narrative radiological interpretations and established visual rating scales (VRSs), calculating the relationship between the Torrison classification and three VRSs: the Medial Temporal Atrophy (MTA) scale, the Koedam scale for posterior atrophy, and the Pasquier scale for Global Cortical Atrophy.

Agreement between categorical or ordinal variables was evaluated using Cohen's kappa (κ) coefficient, while intraclass correlation coefficients (ICCs) were used to assess the consistency of continuous or quasi-continuous variables across methods. Kappa values were interpreted using standard benchmarks: <0.20 (slight), 0.21–0.40 (fair), 0.41–0.60 (moderate), 0.61–0.80 (substantial), and >0.80 (almost perfect agreement).

For each atrophy assessment method (report, VRS, software), concordance rates were calculated to determine how often the methods reached the same classification (normal vs. abnormal). Discrepancies were further analyzed descriptively to explore common patterns of under- or over-reporting.

A p-value < 0.05 was considered statistically significant for all comparisons.

3. Results

3.1. Report - VRS Comparison

Spearman rank correlation analysis demonstrated a moderate to strong positive correlation between the Torrison classification of narrative radiological reports and visual rating scales (VRSs). Specifically, correlation coefficients (ρ) were 0.56 for the Medial Temporal Atrophy (MTA) scale ($p < 0.001$), 0.69 for the Koedam scale ($p < 0.001$), and 0.55 for the Pasquier scale ($p < 0.001$),

For the Pasquier scale, the ICC for individual measures (i.e., comparing each case independently) was 0.54 with a 95% confidence interval ranging from 0.31 to 0.72. This indicates a moderate level of reliability between the visual scale and the radiology report. When assessing the average of the scores across observers or cases, the ICC increased to 0.71 (95% CI: 0.47–0.84), reflecting improved consistency. However, when using Cohen's kappa, which looks at categorical agreement, the value was 0.21 (95% CI: 0.01–0.41), which corresponds to only modest agreement between the two reporting approaches.

A similar trend was observed with the Koedam scale. The ICC for single measures was 0.55 (95% CI: 0.078–0.78), again indicating a moderate level of consistency. The ICC for average measures was slightly higher at 0.71 (95% CI: 0.14–0.87). The Cohen's kappa coefficient for the Koedam scale and narrative reports was 0.30 (95% CI: 0.14–0.46), indicating fair agreement.

When evaluating the MTA scale, the ICC for single scores was 0.47 (95% CI: 0.23–0.72), with the mean score ICC reaching 0.68 (95% CI: 0.18–0.78). This shows that, while there is some alignment between narrative reporting and standardized visual evaluation, the agreement is only moderate at best.

The ICC results are reported in Table 2.

Table 2. Agreement between radiology reports and visual rating scales (VRSs): intraclass correlation coefficients.

Visual Rating Scale	ICC (Single Measures)	ICC (Average Measures)
Pasquier (GCA)	0.54 (95% CI: 0.31–0.72)	0.70 (95% CI: 0.47–0.83)
Koedam	0.55 (95% CI: 0.07–0.78)	0.71 (95% CI: 0.14–0.87)
MTA (Scheltens)	0.47 (95% CI: 0.23–0.72)	0.68 (95% CI: 0.18–0.78)

3.2. VRS - Software Comparison

The mean score derived from the visual scales was 0.196, whereas the software yielded a higher average score of 0.279. This suggests that, on average, visual assessment tends to underestimate the degree of brain atrophy compared to the objective quantitative values generated by the software.

For the MTA scale, the kappa was 0.14, with 71.73% concordance and a statistically significant McNemar p-value of 0.005, suggesting a discrepancy in how often the software and scale classified the same cases as abnormal.

For the Koedam scale, the kappa was 0.33, with 82.61% concordance and a non-significant McNemar $p = 0.28$.

For the Pasquier scale, the kappa was 0.29, concordance was 80.43%, and the McNemar $p = 0.18$.

Concordance results between visual rating scales and software-based volumetric analysis are summarized in Table 3.

Table 3. Comparison of agreement between visual rating scales and the software-based volumetry.

VRS—Software	Cohen Kappa Test	Concordance	McNemar Test—*p* Value
MTA	0.14	71.73%	0.005
Koedam	0.33	82.61%	0.28
Pasquier	0.29	80.43%	0.18

3.3. Report—Software Comparison

Looking at agreement between software and radiologist descriptions for temporal, global, and posterior brain regions, in the temporal region, the kappa was 0.29, with a concordance rate of 84.78%, and a McNemar $p = 0.13$.

For the global cortical region, agreement was markedly lower: kappa = 0.03, concordance = 69.56%, and McNemar $p = 0.50$.

In the posterior region, the kappa was 0.20, with 80.43% concordance and a McNemar $p = 0.78$.

It is important to note that the McNemar test values are non-significant across temporal ($p = 0.13$), global ($p = 0.50$), and posterior ($p = 0.78$) regions. This suggests that the discrepancies observed between the qualitative radiological reports and the quantitative software measurements may not be statistically significant.

Results of the concordance evaluation between radiology reports and automated volumetric measurements are summarized in Table 4.

Table 4. Comparison of the agreement between radiological reports and software-based volumetric analysis.

Report—Software	Cohen Kappa	Concordance	McNemar Test—*p* Value
Temporal	0.29	84.78%	0.13
Global	0.03	69.56%	0.50
Posterior	0.20	80.43%	0.78

To summarize the overall diagnostic patterns, we dichotomized the findings from each method into “normal” (0) and “pathological” (1) and directly compared the frequency of classifications across the three assessment modalities:

- Narrative radiology reports classified 17 cases as pathological and 26 as normal.
- Visual rating scales reported only 8 cases as pathological and 35 as normal.
- Software-based volumetry identified 12 cases as pathological and 31 as normal.

This comparison, illustrated in Figure 2, highlights a key trend: visual rating scales appear to under-report pathological atrophy, while radiology reports tend to overestimate it compared to the more conservative and statistically grounded classifications provided by the software.

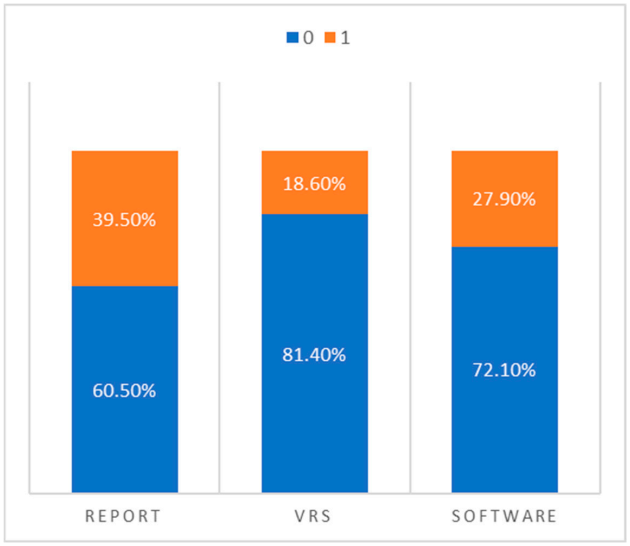

Figure 2. Percentage distribution of "normal" vs. "pathological" classifications across radiology reports, visual rating scales, and automated volumetric software.

4. Discussion

Accurately detecting and characterizing brain atrophy is a central goal in the diagnostic work-up of brain aging, dementia syndromes and several neurodegenerative disorders. A variety of assessment tools exist—ranging from traditional narrative reports to visual rating scales (VRSs) and fully automated volumetric software—each with distinct advantages and limitations [39].

Our study compared these three approaches in a real-world cohort, revealing important discrepancies in classification, agreement, and reliability.

Narrative reports, while dominant in routine clinical practice, rely heavily on the subjective impression of the radiologist, informed by clinical context and visual inspection of MRI or CT scans. Despite their ubiquity, they are limited by low sensitivity and high inter-observer variability. As Harper et al. noted, narrative reporting often fails to detect subtle atrophic changes and is unsuitable for longitudinal monitoring, largely due to the absence of standardized descriptors. Even among experienced neuroradiologists, inter-rater agreement for identifying hippocampal atrophy or posterior cortical thinning has been shown to be poor [14,40].

In our study, although narrative reports moderately correlated with structured VRS assessments (ρ = 0.56–0.69), the agreement was only modest when categorized, as evidenced by low kappa values (0.21–0.30). This reflects the qualitative, variable nature of narrative descriptions and the lack of standard reporting frameworks in current practice. Furthermore, narrative reports tended to overestimate pathological atrophy relative to both VRS and volumetric analysis, likely reflecting an interpretive bias toward caution.

Semi-quantitative visual rating scales represent a compromise between subjective and quantitative methods. Commonly used tools include the Medial Temporal Atrophy (MTA), Global Cortical Atrophy (GCA), and Posterior Atrophy (Koedam) scales. These tools have been widely validated in dementia research and show moderate to good correlation with neuropathological and clinical severity markers. For example, Fumagalli et al. demonstrated that a visual scale specifically assessing parieto-occipital sulcus widening could reliably distinguish posterior cortical atrophy (PCA) from typical Alzheimer's disease, highlighting the value of tailored scales in subtype differentiation [4,8,22].

Our results align with these observations. The MTA scale, in particular, showed a statistically significant association with hippocampal volumes from the Pixyl software and had the highest clinical consistency. However, agreement with software remained low

across all scales when analyzed categorically (e.g., κ = 0.14–0.33), and VRSs consistently underestimated atrophy, especially in borderline cases. As Harper et al. also noted, while VRSs improve diagnostic consistency over narrative reporting, they are still affected by inter-rater variability and require considerable experience for reliable use [14].

Automated volumetric tools, such as commercially available AI based software, offer objective, reproducible brain volume assessments using normative databases adjusted for age and sex. This facilitates early detection of neurodegeneration and is particularly useful in ambiguous or early-stage cases. In line with our findings, Zilioli et al. reported that volumetric analysis better detected early hippocampal and parietal atrophy and more accurately predicted amyloid PET positivity than visual assessment in patients being evaluated for anti-amyloid therapy. They found automated tools to be especially valuable in identifying candidates for disease-modifying treatments [22].

Persson et al. also demonstrated a strong correlation (r = 0.79) between NeuroQuant-derived hippocampal volumes and MTA scores, with volumetry better differentiating MCI from early dementia. They emphasized that automated quantification adds diagnostic value in borderline presentations—supporting our own observation that narrative reports and VRSs diverge most notably from software results in mild or ambiguous cases [21].

Despite their promise, volumetric tools are not without limitations. One key concern is inter-software variability. As Pemberton et al. highlighted in a systematic review, different commercial software platforms may yield non-interchangeable results due to differences in segmentation algorithms, reference datasets, and preprocessing pipelines [41]. This variability complicates multicenter comparisons and reduces confidence in cross-software reproducibility. Our findings also reflect this criticism: while Pixyl provides useful normative comparisons, discrepancies with VRSs and narrative interpretations remain significant and unresolved.

Additionally, our study has limitations. The sample size was relatively small, and many patients had little or no measurable atrophy, limiting statistical power. It is plausible that agreement across methods would improve in cohorts with more advanced neurodegeneration. The Torisson system used to retrospectively score narrative reports is also not a validated clinical tool but provides a useful framework for qualitative-to-quantitative comparison. A key limitation of this study is the absence of correlation with clinical cognitive scales and other imaging modalities, which would be essential for comprehensive validation of imaging findings.

This study demonstrates that while narrative reports remain the standard in clinical neuroimaging, they frequently diverge from structured visual scales and automated software tools—especially in early or subtle presentations of brain atrophy. Visual rating scales offer better standardization but still may underestimate atrophy. Automated volumetric software provides greater sensitivity and reproducibility, especially in ambiguous cases, but suffers from tool-dependent variability that limits comparability.

Future efforts should focus on standardizing VRS use in clinical reports through training and templates, improving volumetric software harmonization across platforms, and developing integrated reporting models that combine qualitative and quantitative findings to enhance diagnostic confidence and accuracy.

Author Contributions: Conceptualization, F.B. and C.F.; methodology, G.S.; software, F.B. and A.I.; validation, C.F., G.D.C. and G.S.; formal analysis, C.D.F.; investigation, C.F.; resources, F.B.; data curation, A.S. (Alessandra Sabatelli) and G.S.; writing—original draft preparation, A.S. (Alessandra Sabatelli) and C.F.; writing—review and editing, G.S.; visualization, F.B.; supervision, F.B. and E.D.C.; project administration, A.S. (Alessandra Splendiani); funding acquisition, A.S. (Alessandra Splendiani). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Our study received approval from the Internal Review Board of the University of L'Aquila (protocol code 21 January 2020 n. 01/2020), and all participating patients provided signed informed consent to take part in the study.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The datasets generated and analyzed for the current study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Agren, R.; Awad, A.; Blomstedt, P.; Fytagoridis, A. Voxel-Based Morphometry of Cerebellar Lobules in Essential Tremor. *Front. Aging Neurosci.* **2021**, *13*, 667854. [CrossRef] [PubMed]
2. Amiri, H.; de Sitter, A.; Bendfeldt, K.; Battaglini, M.; Gandini Wheeler-Kingshott, C.A.M.; Calabrese, M.; Geurts, J.J.G.; Rocca, M.A.; Sastre-Garriga, J.; Enzinger, C.; et al. Urgent challenges in quantification and interpretation of brain grey matter atrophy in individual MS patients using MRI. *Neuroimage Clin.* **2018**, *19*, 466–475. [CrossRef] [PubMed]
3. Cameron, E.; Dyke, J.P.; Hernandez, N.; Louis, E.D.; Dydak, U. Cerebral gray matter volume losses in essential tremor: A case-control study using high resolution tissue probability maps. *Park. Relat. Disord.* **2018**, *51*, 85–90. [CrossRef]
4. Torisson, G.; van Westen, D.; Stavenow, L.; Minthon, L.; Londos, E. Medial temporal lobe atrophy is underreported and may have important clinical correlates in medical inpatients. *BMC Geriatr.* **2015**, *15*, 65. [CrossRef]
5. Zivadinov, R.; Bergsland, N.; Korn, J.R.; Dwyer, M.G.; Khan, N.; Medin, J.; Price, J.C.; Weinstock-Guttman, B.; Silva, D.; Group, M.-M.S. Feasibility of Brain Atrophy Measurement in Clinical Routine without Prior Standardization of the MRI Protocol: Results from MS-MRIUS, a Longitudinal Observational, Multicenter Real-World Outcome Study in Patients with Relapsing-Remitting MS. *AJNR Am. J. Neuroradiol.* **2018**, *39*, 289–295. [CrossRef]
6. Grisoli, M.; Nigri, A.; Medina Carrion, J.P.; Palermo, S.; Demichelis, G.; Giacosa, C.; Mongelli, A.; Fichera, M.; Nanetti, L.; Mariotti, C. Tracking longitudinal thalamic volume changes during early stages of SCA1 and SCA2. *Radiol. Med.* **2024**, *129*, 1215–1223. [CrossRef]
7. Giorgio, A.; De Stefano, N. Advanced Structural and Functional Brain MRI in Multiple Sclerosis. *Semin. Neurol.* **2016**, *36*, 163–176. [CrossRef]
8. Kaushik, S.; Vani, K.; Chumber, S.; Anand, K.S.; Dhamija, R.K. Evaluation of MR Visual Rating Scales in Major Forms of Dementia. *J Neurosci Rural Pr.* **2021**, *12*, 16–23. [CrossRef]
9. Pizzini, F.B.; Conti, E.; Bianchetti, A.; Splendiani, A.; Fusco, D.; Caranci, F.; Bozzao, A.; Landi, F.; Gandolfo, N.; Farina, L.; et al. Radiological assessment of dementia: The Italian inter-society consensus for a practical and clinically oriented guide to image acquisition, evaluation, and reporting. *Radiol. Medica* **2022**, *127*, 998–1022. [CrossRef]
10. Rocca, M.A.; Battaglini, M.; Benedict, R.H.; De Stefano, N.; Geurts, J.J.; Henry, R.G.; Horsfield, M.A.; Jenkinson, M.; Pagani, E.; Filippi, M. Brain MRI atrophy quantification in MS: From methods to clinical application. *Neurology* **2017**, *88*, 403–413. [CrossRef]
11. Sinnecker, T.; Schadelin, S.; Benkert, P.; Ruberte, E.; Amann, M.; Lieb, J.M.; Naegelin, Y.; Muller, J.; Kuhle, J.; Derfuss, T.; et al. Brain atrophy measurement over a MRI scanner change in multiple sclerosis. *Neuroimage Clin.* **2022**, *36*, 103148. [CrossRef] [PubMed]
12. Pizzini, F.B.; Boscolo Galazzo, I.; Natale, V.; Ribaldi, F.; Scheffler, M.; Caranci, F.; Lovblad, K.O.; Menegaz, G.; Frisoni, G.B.; Gunther, M. Insights into single-timepoint ASL hemodynamics: What visual assessment and spatial coefficient of variation can tell. *Radiol. Med.* **2024**, *129*, 467–477. [CrossRef] [PubMed]
13. Hakansson, C.; Torisson, G.; Londos, E.; Hansson, O.; Bjorkman-Burtscher, I.M.; van Westen, D. Reporting frequency of radiology findings increases after introducing visual rating scales in the primary care diagnostic work up of subjective and mild cognitive impairment. *Eur. Radiol.* **2021**, *31*, 666–673. [CrossRef] [PubMed]
14. Harper, L.; Barkhof, F.; Fox, N.C.; Schott, J.M. Using visual rating to diagnose dementia: A critical evaluation of MRI atrophy scales. *J. Neurol. Neurosurg. Psychiatry* **2015**, *86*, 1225–1233. [CrossRef]
15. Kim, G.H.; Kim, J.E.; Choi, K.G.; Lim, S.M.; Lee, J.M.; Na, D.L.; Jeong, J.H. T1-weighted axial visual rating scale for an assessment of medial temporal atrophy in Alzheimer's disease. *J. Alzheimers Dis.* **2014**, *41*, 169–178. [CrossRef]
16. Loreto, F.; Gontsarova, A.; Scott, G.; Patel, N.; Win, Z.; Carswell, C.; Perry, R.; Malhotra, P. Visual atrophy rating scales and amyloid PET status in an Alzheimer's disease clinical cohort. *Ann. Clin. Transl. Neurol.* **2023**, *10*, 619–631. [CrossRef]
17. Mossa-Basha, M.; Andre, J.B.; Yuh, E.; Hunt, D.; LaPiana, N.; Howlett, B.; Krakauer, C.; Crane, P.; Nelson, J.; DeZelar, M.; et al. Comparison of brain imaging and physical health between research and clinical neuroimaging cohorts of ageing. *Br. J. Radiol.* **2024**, *97*, 614–621. [CrossRef]

18. Zhang, Y.; Tartaglia, M.C.; Schuff, N.; Chiang, G.C.; Ching, C.; Rosen, H.J.; Gorno-Tempini, M.L.; Miller, B.L.; Weiner, M.W. MRI signatures of brain macrostructural atrophy and microstructural degradation in frontotemporal lobar degeneration subtypes. *J. Alzheimers Dis.* **2013**, *33*, 431–444. [CrossRef]
19. Brusini, I.; MacNicol, E.; Kim, E.; Smedby, O.; Wang, C.; Westman, E.; Veronese, M.; Turkheimer, F.; Cash, D. MRI-derived brain age as a biomarker of ageing in rats: Validation using a healthy lifestyle intervention. *Neurobiol. Aging* **2022**, *109*, 204–215. [CrossRef]
20. Hofmann, S.M.; Beyer, F.; Lapuschkin, S.; Goltermann, O.; Loeffler, M.; Muller, K.R.; Villringer, A.; Samek, W.; Witte, A.V. Towards the interpretability of deep learning models for multi-modal neuroimaging: Finding structural changes of the ageing brain. *Neuroimage* **2022**, *261*, 119504. [CrossRef]
21. Persson, K.; Barca, M.L.; Edwin, T.H.; Cavallin-Eklund, L.; Tangen, G.G.; Rhodius-Meester, H.F.M.; Selbaek, G.; Knapskog, A.B.; Engedal, K. Regional MRI volumetry using NeuroQuant versus visual rating scales in patients with cognitive impairment and dementia. *Brain Behav.* **2024**, *14*, e3397. [CrossRef]
22. Zilioli, A.; Rosenberg, A.; Mohanty, R.; Matton, A.; Granberg, T.; Hagman, G.; Lotjonen, J.; Kivipelto, M.; Westman, E. Brain MRI volumetry and atrophy rating scales as predictors of amyloid status and eligibility for anti-amyloid treatment in a real-world memory clinic setting. *J. Neurol.* **2024**, *272*, 84. [CrossRef] [PubMed]
23. Vosshenrich, J.; Brantner, P.; Cyriac, J.; Jadczak, A.; Lieb, J.M.; Blackham, K.A.; Heye, T. Quantifying the Effects of Structured Reporting on Report Turnaround Times and Proofreading Workload in Neuroradiology. *Acad. Radiol.* **2023**, *30*, 727–736. [CrossRef]
24. Goodkin, O.; Pemberton, H.; Vos, S.B.; Prados, F.; Sudre, C.H.; Moggridge, J.; Cardoso, M.J.; Ourselin, S.; Bisdas, S.; White, M.; et al. The quantitative neuroradiology initiative framework: Application to dementia. *Br. J. Radiol.* **2019**, *92*, 20190365. [CrossRef]
25. Mallio, C.A.; Sertorio, A.C.; Bernetti, C.; Beomonte Zobel, B. Large language models for structured reporting in radiology: Performance of GPT-4, ChatGPT-3.5, Perplexity and Bing. *Radiol. Med.* **2023**, *128*, 808–812. [CrossRef]
26. Mallio, C.A.; Bernetti, C.; Sertorio, A.C.; Beomonte Zobel, B. Large language models and structured reporting: Never stop chasing critical thinking. *Radiol. Med.* **2023**, *128*, 1445–1446. [CrossRef]
27. Mamlouk, M.D.; Chang, P.C.; Saket, R.R. Contextual Radiology Reporting: A New Approach to Neuroradiology Structured Templates. *AJNR Am. J. Neuroradiol.* **2018**, *39*, 1406–1414. [CrossRef]
28. Granata, V.; Fusco, R.; Coluccino, S.; Russo, C.; Grassi, F.; Tortora, F.; Conforti, R.; Caranci, F. Preliminary data on artificial intelligence tool in magnetic resonance imaging assessment of degenerative pathologies of lumbar spine. *Radiol. Med.* **2024**, *129*, 623–630. [CrossRef]
29. Lee, J.; Lee, J.Y.; Oh, S.W.; Chung, M.S.; Park, J.E.; Moon, Y.; Jeon, H.J.; Moon, W.J. Evaluation of Reproducibility of Brain Volumetry between Commercial Software, Inbrain and Established Research Purpose Method, FreeSurfer. *J. Clin. Neurol.* **2021**, *17*, 307–316. [CrossRef]
30. Kim, S.H.; Schramm, S.; Riedel, E.O.; Schmitzer, L.; Rosenkranz, E.; Kertels, O.; Bodden, J.; Paprottka, K.; Sepp, D.; Renz, M.; et al. Automation bias in AI-assisted detection of cerebral aneurysms on time-of-flight MR angiography. *Radiol. Med.* **2025**, *130*, 555–566. [CrossRef]
31. Cover, K.S.; van Schijndel, R.A.; van Dijk, B.W.; Redolfi, A.; Knol, D.L.; Frisoni, G.B.; Barkhof, F.; Vrenken, H.; neuGrid; Alzheimer's Disease Neuroimaging, Initiative. Assessing the reproducibility of the SienaX and Siena brain atrophy measures using the ADNI back-to-back MP-RAGE MRI scans. *Psychiatry Res.* **2011**, *193*, 182–190. [CrossRef] [PubMed]
32. Koussis, P.; Toulas, P.; Glotsos, D.; Lamprou, E.; Kehagias, D.; Lavdas, E. Reliability of automated brain volumetric analysis: A test by comparing NeuroQuant and volBrain software. *Brain Behav.* **2023**, *13*, e3320. [CrossRef]
33. Lee, J.Y.; Park, J.E.; Chung, M.S.; Oh, S.W.; Moon, W.J. [Expert Opinions and Recommendations for the Clinical Use of Quantitative Analysis Software for MRI-Based Brain Volumetry]. *Taehan Yongsang Uihakhoe Chi* **2021**, *82*, 1124–1139. [CrossRef]
34. Tanabe, J.; Lim, M.F.; Dash, S.; Pattee, J.; Steach, B.; Pressman, P.; Bettcher, B.M.; Honce, J.M.; Potigailo, V.A.; Colantoni, W.; et al. Automated Volumetric Software in Dementia: Help or Hindrance to the Neuroradiologist? *AJNR Am. J. Neuroradiol.* **2024**, *45*, 1737–1744. [CrossRef]
35. Bruno, F.; Tommasino, E.; Catalucci, A.; Pastorelli, C.; Borea, F.; Caldarelli, G.; Bellini, M.; Badini, P.; Mancini, S.; Santobuono, C.; et al. Evaluation of Cerebral Volume Changes in Patients with Tremor Treated by MRgFUS Thalamotomy. *Life* **2022**, *13*, 16. [CrossRef]
36. Scheltens, P.; Leys, D.; Barkhof, F.; Huglo, D.; Weinstein, H.C.; Vermersch, P.; Kuiper, M.; Steinling, M.; Wolters, E.C.; Valk, J. Atrophy of medial temporal lobes on MRI in "probable" Alzheimer's disease and normal ageing: Diagnostic value and neuropsychological correlates. *J. Neurol. Neurosurg. Psychiatry* **1992**, *55*, 967–972. [CrossRef]
37. Pasquier, F.; Leys, D.; Weerts, J.G.; Mounier-Vehier, F.; Barkhof, F.; Scheltens, P. Inter- and intraobserver reproducibility of cerebral atrophy assessment on MRI scans with hemispheric infarcts. *Eur. Neurol.* **1996**, *36*, 268–272. [CrossRef]
38. Koedam, E.L.; Lehmann, M.; van der Flier, W.M.; Scheltens, P.; Pijnenburg, Y.A.; Fox, N.; Barkhof, F.; Wattjes, M.P. Visual assessment of posterior atrophy development of a MRI rating scale. *Eur. Radiol.* **2011**, *21*, 2618–2625. [CrossRef]

39. Cole, J.H.; Franke, K. Predicting Age Using Neuroimaging: Innovative Brain Ageing Biomarkers. *Trends Neurosci.* **2017**, *40*, 681–690. [CrossRef]
40. Parillo, M.; Vaccarino, F.; Beomonte Zobel, B.; Mallio, C.A. ChatGPT and radiology report: Potential applications and limitations. *Radiol. Med.* **2024**, *129*, 1849–1863. [CrossRef]
41. Pemberton, H.G.; Zaki, L.A.M.; Goodkin, O.; Das, R.K.; Steketee, R.M.E.; Barkhof, F.; Vernooij, M.W. Technical and clinical validation of commercial automated volumetric MRI tools for dementia diagnosis-a systematic review. *Neuroradiology* **2021**, *63*, 1773–1789. [CrossRef] [PubMed]