types [5]. These systematic and random variations are machine dependent and can be corrected for the most part via image denoising [6], bias field inhomogeneity estimation [7], and intensity standardization [8].

In this paper we investigated the reliability of our DEF metric through analysis of cross-sectional (i.e., one timepoint) scan/repeat scan and scan/rescan images from two multicentric studies. First, we took advantage of the fact that subjects in the Alzheimer's Disease Neuroimaging Initiative (ADNI) study received two within-session T1-weighted scans at their baseline visit to test for scan/repeat scan analysis. Further, we employed data on three participants in the Pilot European ADNI that had been scanned at seven different sites in a short timeframe to test for Scan/Rescan reliability. We report minimum clinical trial sample size increases at various different levels based on the calculated detection threshold.

Reliability analysis is an important, necessary, and often overlooked step between bench and bedside in the research and clinical contexts.

2. Materials and Methods

2.1. Ethics. Institutional review boards of all participating institutions approved the procedures for this study. Written informed consent was obtained from all participants or surrogates. More information about ADNI[1] and Pilot European ADNI investigators are provided in the Acknowledgments.

2.2. Subjects. In this study we used data from three different studies, totaling 1051 subjects from over 60 centers.

(i) The first was the *Mapping group*, consisting of 145 young control subjects from the International Consortium for Brain Mapping database [9].

(ii) The second was the *Classification group*, which consisted in 70 probable AD and 69 CTRL subjects from the LENITEM database [10]. We required those first two groups to build our high-dimensional metric;

(iii) The third was the *Scan/Repeat Test Group*, which consisted in 1518 baseline MRIs (scan + same-session repeat scans) from 759 CTRL, MCI, and probable AD subjects participating in ADNI, acquired on more than 50 different 1.5T scanners using a similar 3D T1-weighted MP-RAGE protocol [11]. Inclusion criteria to the ADNI study were as follows.

(a) CTRL are MMSE scores [12] between 24–30 (inclusive), a CDR [13] of 0, nondepressed, non-MCI, and nondemented. The age range of normal subjects was roughly matched to that of MCI and mild AD subjects.

(b) MCI subjects are MMSE scores between 24–30 (inclusive), a memory complaint, objective memory loss measured by education adjusted scores on Wechsler Memory Scale Logical Memory II [14], a CDR of 0.5, absence of significant levels of impairment in other cognitive domains, essentially preserved activities of daily living, and an absence of dementia.

(c) Mild AD is MMSE scores between 20–26 (inclusive), CDR of 0.5 or 1.0, and meets NINCDS/ADRDA criteria for probable AD [15].

From the complete ADNI dataset of 822 subjects at baseline, we selected individuals for the *Scan/Repeat Test Group* that had both valid entry images and processed images that passed *automated* quality control [16].

(iv) Finally, the fourth was the *Scan/Rescan Test Group*, which was obtained with permission from the multicentric Pilot European ADNI project [17]. It included data from three healthy volunteers acting as human quality control phantoms for the study.

2.3. MRI Acquisitions. Subjects in the *Mapping group* were scanned in Montreal, QC, Canada on a Philips Healthcare Gyroscan 1.5T scanner (Best, The Netherlands) using a T1-weighted fast gradient echo sequence (sagittal acquisition, TR = 18 ms, TE = 10 ms, $1 \times 1 \times 1\,\text{mm}^3$ voxels, flip angle 30°).

Subjects in the *Classification group* were scanned in Brescia, Italy on a single Philips Healthcare Gyroscan 1.0T scanner (Best, The Netherlands) using a T1-weighted fast field echo sequence (sagittal acquisition, TR = 25 ms, TE = 6.9 ms, $1 \times 1 \times 1,3\,\text{mm}^3$ voxels).

Subjects in the *Scan/Repeat Test Group* were scanned on over 50 different 1.5T scanners (GE Medical Systems; Siemens Healthcare; Philips Healthcare) using a 3D T1-weighted MP-RAGE protocol or its equivalent [11]. In this protocol, within the same scan session, there were two 3D T1-weighted images acquired, allowing us to test reliability on this scan/repeat pair. The subject was not taken out of the scanner between acquisitions.

Subjects in the *Scan/Rescan Test Group* were scanned within the span of few weeks at seven different European centers (Sites 1 to 7), using the ADNI study 3D T1-weighted MP-RAGE protocol [11]. Six centers collected scan/rescan sessions, where the subject was taken out of the scanner between acquisitions. This allows us to estimate scan/rescan reliability on 18 comparison pairs.

2.4. Initial Image Processing. We processed all MRI volumes identically using the MINC image processing toolbox (http://www.bic.mni.mcgill.ca/ServicesSoftware/HomePage) and local software as follows: (a) noise removal [6]; (b) raw scanner intensity inhomogeneity correction [7]; (c) global registration (12 degrees of freedom) [18] to the reference image space defined by the BrainWeb T1-weighted image [19] (1-mm resolution, 0% noise, 0% nonuniformity), maximizing the mutual information between the two volumes [20]; (d) resampling to a 1-mm^3 isotropic grid; (e) linear clamping to (0–100) intensity range; (f) intensity standardization [8]; (g) nonlinear registration of individual standardized subject images to the BrainWeb reference;