FreeSurfer-based segmentation of hippocampal subfields: A review of methods and applications, with a novel quality control procedure for ENIGMA studies and other collaborative efforts

Philipp G. Sämann1 | Juan Eugenio Iglesias2,3,4 | Boris Gutman5 | Dominik Grotegerd6 | Ramona Leenings6 | Claas Flint6,7 | Udo Dannlowski6 | Emily K. Clarke-Rubright8,9 | Rajendra A. Morey8,9 | Theo G.M. van Erp10,11 | Christopher D. Whelan12 | Laura K. M. Han13 | Laura S. van Velzen14,15 | Bo Cao16 | Jean C. Augustinack3 | Paul M. Thompson12 | Neda Jahanshad12 | Lianne Schmaal14,15

1Max Planck Institute of Psychiatry, Munich, Germany
2Centre for Medical Image Computing, University College London, London, UK
3The Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital/Harvard Medical School, Boston, Massachusetts
4Computer Science and AI Laboratory (CSAIL), Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts
5Department of Biomedical Engineering, Illinois Institute of Technology, Chicago
6Department of Psychiatry, University of Münster, Münster, Germany
7Department of Mathematics and Computer Science, University of Münster, Germany
8Brain Imaging and Analysis Center, Duke University, Durham, North Carolina
9VISN 6 MIRECC, Durham VA, Durham, North Carolina
10Clinical Translational Neuroscience Laboratory, Department of Psychiatry and Human Behavior, University of California, Irvine, California
11Center for the Neurobiology of Learning and Memory, University of California Irvine, Irvine, California
12Imaging Genetics Center, Stevens Neuroimaging and Informatics Institute, Keck School of Medicine, University of Southern California, Los Angeles, California
13Department of Psychiatry, Amsterdam University Medical Centers, Vrije Universiteit and GGZ inGeest, Amsterdam Neuroscience, Amsterdam, The Netherlands
14Orygen, Parkville, Australia
15Centre for Youth Mental Health, The University of Melbourne, Melbourne, Australia
16Department of Psychiatry, Faculty of Medicine & Dentistry, University of Alberta, Edmonton, Canada

[Corrections added in affiliations on 02nd January, 2021, after first online publication]

Correspondence
Philipp G. Sämann, Max Planck Institute of Psychiatry, Kraepelinstr. 2-10, 80804 Munich, Germany.
Email: saemann@psych.mpg.de

Funding information
European Research Council; National Institutes of Health, Grant/Award Number: U54 EB020403; Interdisciplinary Center for Clinical Neurosciences

Abstract
Structural hippocampal abnormalities are common in many neurological and psychiatric disorders, and variation in hippocampal measures is related to cognitive performance and other complex phenotypes such as stress sensitivity. Hippocampal subregions are increasingly studied, as automated algorithms have become available for mapping and volume quantification. In the context of the Enhancing Neuro Imaging Genetics Initiative (ENIGMA), a collaborative effort for the sharing of large-scale multimodal brain image and clinical data, a novel quality control procedure is described, which is crucial for main analyses and meta-analyses. Detailed guidelines are provided, and the toolkit is open-access to facilitate further work in this field.
1 | INTRODUCTION

The hippocampus is an intensely studied brain region in preclinical and clinical neuroscience (Andersen, Morris, Amaral, Bliss, & O’Keefe, 2007). It is involved in multiple aspects of cognition, including spatial navigation (O’Keefe, 1990), learning (Morris, 2006), and episodic memory (Burgess, Maguire, & O’Keefe, 2002), affective processing (Koelsch et al., 2015), and stress response regulation in humans (Herman et al., 2016; McEwen & Akil, 2020). Given its many functions—and its circumscribed, exposed position in the medial temporal lobe—it has been the target of volumetric studies in neurology and psychiatry. Early hippocampal studies with computed tomography used indirect metrics such as widened hippocampal fissure, increased hippocampal lucency, or temporal horn diameter (de Leon, George, Stylopoulos, Smith, & Miller, 1989; George et al., 1990; Sandor, Albert, Stafford, & Harpley, 1988; Scheltens, Weinstein, & Leys, 1992). Its contrast-rich boundaries on magnetic resonance imaging (MRI) against the ventricular space may even have led to an “investigation bias” toward this structure compared to other less circumscribed brain structures. Meanwhile, MRI-based hippocampal morphometry has advanced through improved contrast, spatial resolution, and computational progress. Images with 0.5–0.9 mm isotropic spatial resolution are now regularly collected with standard anatomical sequences on 3 Tesla platforms (e.g., Human Connectome Project, http://protocols.humanconnectome.org). The spread of high-field MRI further accelerated this development. Recent work validating in vivo 7-Tesla human MRI with histology has pushed MRI toward “in vivo neuropathology” (DeKraker, Ferko, Lau, Köhler, & Khan, 2018) and allows subfield segmentation techniques to take advantage of the resolution and contrast available at 7 Tesla (Giuliano et al., 2017).

Attempts to subdivide the hippocampal formation into substructures from MRI have been made for several years. Based on sublayers and boundaries with neighboring structures, subdividing the hippocampus has been the objective of several manual segmentation protocols (Jeukens et al., 2009; La Joie et al., 2010; Mueller et al., 2007; Mueller et al., 2018; Mueller & Weiner, 2009; Shing et al., 2011; Wisse et al., 2012). Detecting and following boundaries in the complex, “swiss roll”–like structure and recognizing landmarks when the internal contrast is insufficient, is a complex visual task that requires expert anatomical knowledge and practice. The procedure may take up to 40 hours per case (Winterburn et al., 2013), exhausting realistically available resources, particularly in larger imaging genetics studies, some of which now exceed 50,000 individuals (Satizabal et al., 2019, 2020). Manual segmentations are still represented the gold standard and have been carefully interwoven into fully automated algorithms (Iglesias et al., 2015; Van Leemput et al., 2008, 2009; Yushkevich et al., 2009; Yushkevich et al., 2015).

The Enhancing Neuro Imaging Genetics through Meta Analysis (ENIGMA) consortium conducts imaging and imaging genetics analyses in population cohorts and clinical samples using an international multisite framework, bringing together imaging and genetics groups for large-scale collaborations. The principle of distributed analyses—in which local sites perform image postprocessing and association statistics on their own authority—makes harmonization an important
element to enhance reliability. Collaborative studies allow for detection of small effects sizes, but, the resulting summary statistics rely heavily on the quality of the incoming local association results and extracted data. As disease working groups are increasingly engaged (Thompson et al., 2020) in cross-diagnostic comparisons or genuine transdiagnostic ("denosologization") studies, these efforts might benefit from pipeline harmonization (Figure 1).

The purpose of this report is to give an overview of the hippocampal subfield segmentation algorithm implemented in FreeSurfer (FS) and present a novel protocol for standardizing quality control (QC). To put the pipeline into context and to guide researchers to optimally use the algorithm, we review reports that have employed this tool, pointing out typical methodological challenges and pitfalls. In this article, we follow the recently established practice to refer to the resulting substructures either as "subregion(s)" or "subfield(s)", in an interchangeable way despite the erstwhile anatomical nomenclatures that reserved "subfields" for the cornu ammonis (CA). Section 1 provided background and rationale for this study. In Section 2, we describe the core algorithm of the FS-based hippocampus segmentation, summarize findings of reliability studies, with examples of the correlation between subfield volumes, and discuss the relationship between size and reliability. Next, Section 3 describes the QC procedure developed in the context of the ENIGMA consortium to improve pipeline harmonization. In Section 4, we give an overview on FS-based hippocampal subfield studies published between January 2013 and December 2019, focusing on (a) the clinical or neuroscience domains being studied, (b) statistical methods, and (c) QC procedures. Finally, Section 5 concludes with a summary, discussion, and outlook.

2 | THE FS HIPPOCAMPAL SUBFIELD SEGMENTATION ALGORITHM

2.1 | Principle and output

The hippocampal segmentation tool examined in this paper is part of the widely used and freely available neuroimaging analysis package, FS (Fischl, 2012). The current version of this tool is FreeSurfer 7.1,
and this version has no significant changes to the hippocampal subfield tool compared to version 6.0 [FS6.0]. FreeSurfer 7.1 includes a segmentation of the amygdala (Saygin et al., 2017) that does not affect the hippocampal subregion segmentation that was introduced with FS6.0. The automated segmentation of hippocampal subfields is driven by a probabilistic atlas, built from two manually delineated datasets: one comprised of 15 post mortem subjects, including four with Alzheimer’s disease, from whom ex vivo MRI (7 Tesla, FLASH sequence) was obtained at very high resolution (0.10–0.20 mm isotropic); and a second that comprised 39 T1-weighted in vivo images (1 mm isotropic voxels) from 29 controls and 10 mildly demented individuals. The key developmental leap from FS5.3 to FS6.0 was the inclusion of manually delineated high resolution ex vivo MRI data (Iglesias et al., 2015). Both labeled datasets were combined by a Bayesian algorithm into a probabilistic atlas that is encoded in a tetrahedral mesh, where each node carries probabilistic information on its assignment to subregions (van Leemput, 2009).

The challenge of segmenting a new, unseen hippocampus was then posed as a Bayesian inference problem of maximizing the probability of the segmentation—given the atlas and the input image. Robustness to changes in MRI contrast was achieved by modeling voxel intensities as samples from a Gaussian mixture model conditioned on the hidden segmentation. Parameters of this model were informed by the individual input scan more specifically from five

---

**FIGURE 2** Typical subfield size ranking, overlay in FreeSurfer viewer and exemplary age-volume relationships. (a) Typically raw subfield volumes ranked according to their average size are depicted from a local Max Planck Institute of Psychiatry (MPIP)-based sample of N = 614 subjects (T1WI, FS6.0, mean values and 1 SD). The volume ranking order is extremely robust across other samples including 3 Tesla samples (data not shown). The colored frames point to subregions that underlie ranking violation rules (see Section 2.3 for details). (b) A 3 Tesla example viewed in FS in three corresponding planes (see white cross-hair) with the FS inherent color scheme. The same scheme for the 12 subregions was adopted in the ENIGMA quality control (QC) algorithm. (c) An example of a tendency for nonlinear age effects for the bilateral CA1 region, adjusted for intracranial volume (ICV), sex, diagnosis (major depressive disorder [MDD]/healthy) and site (here coding for coil upgrade related raw image differences). Quadratic or cubic polynomial fits are superior to a linear correlation. (d) The same principle is plotted for the ratio between total gray matter volume and ICV; a less strong nonlinear influence can be read from the fit values. (e) One of three aggregation schemes available in recent development versions of FS, referred to as “FS60,” explaining the mapping between the 12 output labels (which are the labels in the FS6.0 atlas used in this study) and the underlying regions.
intensity compartments: gray matter, white matter, cerebrospinal fluid, alveus, and the molecular layer). Although the alveus and molecular layer resemble white matter, they are modeled separately, given that partial volume effects influence their voxel intensities. The outputs of the hippocampal segmentation are left and right hemisphere images with label assignments for voxels in the hippocampal area (at resolutions of 1-mm isotropic and 0.333-mm isotropic) to either background or one of twelve subregions. These are (sorted by average size): CA1, molecular layer (ML), hippocampal tail, subiculum, presubiculum, granule cell layer of dentate gyrus (GC-ML-DG), CA4, CA3, hippocampal fissure, hippocampus-amygdala-transition-area (HATA), and fimbria (Figure 2a,b). The manual delineation protocol and the grouping of its subregions into 12 output labels are detailed in Iglesias et al. (2015). Alternative aggregations, for example, ones that consider head, body and tail subdivisions, are available online (https://surfer.nmr.mgh.harvard.edu/fswiki/HippocampalSubfieldsAndNucleiOfAmygdala). In this report, we use reaggregation to refer to aggregations deviating from the three standard aggregations.

2.2 Measurement reliability and validity

The output of the automated FS subfield segmentation consists of both the spatial map of a subfield and its absolute volume, and reliability measures have been reported for both, as detailed in Sections 2.2.1 and 2.2.2. For volumes, the difference between measurements (e.g., test–retest measures) can be expressed as percentage error or as intraclass coefficient (ICC). The latter quantifies the consistency of repeated measurements of the same quantity. Different subtypes of the ICC exist, and some are sensitive also to systematic biases of one measurement, a feature that may remain undetected by the Pearson correlation. The spatial reliability of subfield segmentations is often expressed as the Sørensen–Dice index often abbreviated to Dice index; range 0–1; 1 represents perfect overlap of both measurements). The Dice index can be calculated for a single pair of repeatedly measured subregions and is relevant for using the subfield segmentation as regions of interest (ROIs) for other (e.g., functional imaging) experiments.

2.2.1 Test–retest reliability (same scanner)

Several studies have examined test–retest reliability of FS6.0 subfield volumes, all reporting ICC values larger than 0.5 and most around 0.9 (Brown et al., 2020; Elvsåshagen et al., 2016; Quattrini et al., 2020; Whelan et al., 2016; Worker et al., 2018). More specifically, Whelan et al. (2016) included data from four independent cohorts (total N > 1,740), reporting high ICC values (0.70–0.97) for a 3 Tesla sample, and moderate to high for a 4 Tesla sample (0.50–0.89). Another 3-Tesla study applied FS6.0 to data from 22 healthy subjects scanned three times, and 40 patients with AD scanned twice, again reporting high ICC values (>0.9) for 20 of 24 investigated hippocampal subregions, and significantly lower ICC for the hippocampal fissure and fimbria (Worker et al., 2018). In addition, high sensitivity toward AD progression at both 2-year and 6-week follow-up times was detected, using intra-individually optimized co-registration of repeated scans (Worker et al., 2018). In their report on bipolar disorder, Elvsåshagen et al. (2016) reported test–retest ICC over 0.94 for 3-Tesla T1-weighted images selectively for the dentate gyrus/C4A, for within-session repeats (N = 53) and two separate scanning sessions (N = 21). Quattrini et al. (2020) also reported good volume reliability (mean error <5%, ICC > 0.92) and excellent spatial agreement (mean Dice index >0.92) for different 3 Tesla scanner platforms. Another recent report on FS6.0 (Brown et al., 2020) confirmed high test–retest reliability (measured by volume differences, ICC, and Dice overlap) with a similar regional ranking, that is, high reliability for all subregions except for the parasubiculum, hippocampal fissure, and HATA.

The reliability of FS5.3 hippocampal subfield segmentations have also been investigated in 65 subjects scanned at 13 sites. Marizzoni et al. (2015) reported good or very good volumetric (reproducibility errors 2–5%) and spatial reproducibility (Dice indices ~0.9 or >0.9) for all subregions except for the hippocampal fissure and fimbria.

Of note, since FS7.1 (released 04/2020), hippocampal subfield segmentation can be combined with longitudinal input data to optimally assess within-subject changes. This approach reduces measurement noise compared with two separate segmentations, increases test–retest reliability and thus leads to higher sensitivity to longitudinal changes (Iglesias et al., 2016). The two more recent reports strongly confirmed this finding both for within- and across-scanner comparisons, particularly for older subjects (Brown et al., 2020; Quattrini et al., 2020). For repeated measures, this approach is thus clearly recommended.

2.2.2 Reliability across vendor platforms and field-strengths

Whelan et al. (2016) reported on a FS6.0 study comparing T1WI of 1.5 Tesla with 3 Tesla and found high ICC for 11 subregions (0.721–0.915) and a low value for the hippocampal fissure (0.575). Another small trans-platform analysis (1.5 and 3 Tesla, FS5.3, performed on seven children) was included in the report of Tamnes et al. (2014), showing high linear correlations (r values .80–.97) for all subfields except the fimbria. More recently, reliability across vendors (or before and after typical hardware upgrades) has been reported in great detail (Brown et al., 2020; Quattrini et al., 2020). As a general pattern, these studies show that there is higher sensitivity of the spatial reproducibility to MR scanner effects, compared to the subfield volumes. Within-session T1 averages are suggested as one possibility to improve test–retest reliability.

2.2.3 Relationship between average subregion size and measurement reliability

For FS5.3, Marizzoni et al. (2015) reported that specific test–retest measures (reproducibility errors and Dice index) were positively
TABLE 1 Correlation between total HV and subregions, corrected for ICV, age, age-squared, sex, and site, in healthy subjects at 1.5 and 3 Tesla. Sorting is according to the partial correlation values ($r_p$) in descending order per sample. The last two columns denote relative shifts of the volume rank and the correlation rank of the 3 Tesla compared with the 1.5 Tesla sample. Negligible volume rank shifts for the two smallest regions (parasubiculum, HATA) were noted. For the correlation rank, identical ranking was found for 6 subregions, and a minor perturbation—by a maximum of two ranks—for the remaining subregions

<table>
<thead>
<tr>
<th>Region</th>
<th>1.5 Tesla (N = 209)</th>
<th>3 Tesla (N = 211)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Volume rank</td>
<td>Fisher’s r</td>
</tr>
<tr>
<td>Molecular layer</td>
<td>2</td>
<td>0.967</td>
</tr>
<tr>
<td>CA1</td>
<td>1</td>
<td>0.863</td>
</tr>
<tr>
<td>GC-ML-DG</td>
<td>6</td>
<td>0.841</td>
</tr>
<tr>
<td>Subiculum</td>
<td>4</td>
<td>0.815</td>
</tr>
<tr>
<td>CA4</td>
<td>7</td>
<td>0.806</td>
</tr>
<tr>
<td>Presubiculum</td>
<td>5</td>
<td>0.662</td>
</tr>
<tr>
<td>Hippocampal tail</td>
<td>3</td>
<td>0.654</td>
</tr>
<tr>
<td>CA3</td>
<td>8</td>
<td>0.588</td>
</tr>
<tr>
<td>HATA</td>
<td>11</td>
<td>0.557</td>
</tr>
<tr>
<td>Hippocampal fissure</td>
<td>9</td>
<td>0.386</td>
</tr>
<tr>
<td>Parasubiculum</td>
<td>12</td>
<td>0.359</td>
</tr>
<tr>
<td>Fimbria</td>
<td>10</td>
<td>0.201</td>
</tr>
</tbody>
</table>

Abbreviation: ICV, intracranial volume.

associated with regional volumes. In the interest of this report, based on published results (Whelan et al., 2016), Table 2), we correlated the rank of the average subregion size with corresponding ICC values for two FS6.0 samples (unweighted averages across both samples and both hemispheres), and found a high positive rank correlation (Spearman’s rho 0.85). This pattern was further confirmed and extended to the Dice overlap index for FS6.0 by Quattrini et al. (2020) who also concluded that the best test–retest reproducibility was achieved for hippocampal subfields >300 mm$^3$ (see typical size ranking in Figure 2a). This could imply that the power to detect group differences or other clinical correlations may be biased toward larger subregions, and this should be considered when interpreting results.

2.2.4 | Other reliability aspects

T2WI as input

It is known that a T2-weighted image (mostly in the coronal plane, perpendicular to the main axis of the hippocampus), facilitates the identification of the molecular layer and of its boundary with the CA1, which are usually not clearly visible on standard T1-weighted data (Iglesias et al., 2015; Wisse et al., 2014). In this context, Mueller et al. (2018) compared four automated segmentation protocols, both T1WI- and T2WI-based and manual segmentation, particularly with regard to increasing the sensitivity to disease effects. Indeed, effect sizes were higher for methods that included a high resolution T2WI compared to solely T1WI-based methods. Yet, no direct comparison between a single T1WI and combined T1-/T2WI input to version 6.0 was reported. Such a direct comparison is reported in Iglesias et al. (2015), who demonstrated increased sensitivity to MCI/controls differences with the combined input.

FS subfield tool to measure total HV

As the total hippocampal volume (HV) is included in some analyses of subfield data (as a dependent variable, or as a correction variable), a recent study ($N = 664$) validated manually traced HVs against several FS versions (5.2, 5.3, and 6.0) (Schmidt et al., 2018). In brief, automatically measured HV values were systematically larger than manually traced HVs, with a small influence of age and HV itself on this bias. Still, ICC values were high as were measures of spatial overlap. The study also reported detailed failure rates for FS6.0, estimated from 708 cases: The general segmentation failed in 25 cases (3.5%), with six cases (0.8%) showing incorrect orientation and two cases (0.3%) showing insufficient quality or contrast, leaving 17 cases (2.4%) without clearly explained failure. Several other reports focused on the reliability of automated total HV measurement: as concluded from eight reliability measures, FS6.0 subfield-based total HV proved more reliable than other automated algorithms (Khelif, Egorova, et al., 2019; Khelif, Werden, et al., 2019), with acceptable biases. When FS6.0 is applied to children and adolescents, estimation biases compared to manual tracing may appear (Herten, Konrad, Krinzinger, Seitz, & von Polier, 2019; Schoemaker et al., 2016). For the analysis of brains with marked atrophy, the
false segmentation of dura in FS has been identified as a problem for medial temporal lobe studies (Xie et al., 2019).

2.2.5 | Measurement validity

Studies of test–retest reliability properties cannot account for deviation from the ground truth labeling, so, regardless of the study design, these studies are not informative on the actual validity of the subfield measurements. General external validity of the FS-based hippocampal subfield volume measures is rooted in the cytoarchitectonic principle of the ex vivo sample labeling. Further validity aspects are still investigated with a latent assumption that subfield measures hold indirect information on functional or circuit differences with (patho-)physiological relevance. Here, studies on subregional functional connectivity (FC), including intrahippocampal connectivity, or morphological covariance analysis provide relevant contributions (Dalton, McCormick, de Luca, Clark, & Maguire, 2019; Dalton, McCormick, & Maguire, 2019; Ge et al., 2019). An intriguing approach to derive hippocampal subregions is to search for subfield-specific connectivity patterns, instead of hard histological ground truth: purely data-driven, FC-based hippocampal parcellations impressively match manual and—to a lesser degree—automated segmented subfields (Wu et al., 2018). Further, combined anatomical and functional features from 3 Tesla delivered a similar accuracy to 7 Tesla data. So far, no FC studies have been presented that directly incorporate FS6.0-based subfield volumes, or that relate subfield volumes to connectivity measures.

One report critically addressed different aspects of the FS5.3 hippocampal subfield protocol (Wisse et al., 2014):

1. The effect of using low resolution T1WI as exclusive input to the algorithm, as these images might not provide sufficient contrast, for example, of the white matter bands between the dentate gyrus and the CA.
2. A bias in the original parcellation scheme that extrapolated one coronal section to the entire longitudinal axis of the hippocampus (Van Leemput et al., 2008, 2009), leading to overestimations of CA1 and atypical result patterns.
3. The mismatch between high-resolution 3 Tesla T2WI (0.19 mm isometric) on which the algorithm has been developed, and its application to low resolution (1 mm isometric) T1WI.

The second concern, also expressed by de Flores et al. (2015), is directed toward a core element of the algorithm and was addressed by integrating ex vivo data into the atlas. This improved the consistency with manual segmentations (i.e., validity), in terms of the boundaries of CA2/3, presubiculum and the transition between the subiculum and CA1 (Iglesias et al., 2015).

The first and third concerns that have been re-emphasized in a recent comment (Wisse et al., 2020) boil down to the question of whether low resolution T1WI as inputs can lead to sufficiently valid subfield results. For optimal correspondence with the gold standard, good GM/WM contrast is required to detect the white matter bands and the molecular layer. There is currently no clear demonstration that this can be only achieved by adding a T2WI, rather than by a T1WI with high resolution at 0.4–0.5 mm isotropic or 0.4 × 0.4 × 1.0 mm³. At an empirical level, also when this feature is suboptimally present, lower resolution T1WI have detected expected biological effects, for example, with good discrimination between MCI and controls (Iglesias et al., 2015). Still, a methodological rather than clinical validation study is missing that compares manual tracings according to the ex vivo protocol on very high resolution MRI data (e.g., 0.2 mm isotropic) to atlas based segmentations of downsampled images or newly acquired lower resolution images of the same subjects.

2.3 | Cross-correlations among subfield volumes and multiple testing considerations

Due to the explorative character of many hippocampal subfield studies, multiple test correction needs to be performed to control for false positive results. In the following exemplary analyses, we demonstrate how the correlation of (bilaterally added) subfield volumes with each other and with total HV in two samples. The purpose of these analyses is to demonstrate how the correlation structure is influenced by the global correction variable (intracranial volume [ICV] vs. total HV), and how subfield volumes relate to the total HV on a 1.5 Tesla compared to a 3 Tesla platform. Sample A was a 1.5 Tesla major depressive disorder (MDD)/controls sample acquired at the Max Planck Institute of Psychiatry (MPfP) and included in previous ENIGMA studies (Schmaal et al., 2017) (General Electric, Signa/Excite, T1WI, SPGR-3D or spin-echo, voxel size 0.9 × 0.9 × 1.2 mm³, N = 597 [209 controls, 388 MDD patients], age 19–89 years); and Sample B, a 3 Tesla sample of healthy subjects (Münster Cohort, 3T Philips Gyroscan Intera, T1WI, 3D fast gradient echo sequence, voxel size 0.5 mm isometric; N = 211, age 17–61 years). The data were processed with FS5.3 (general segmentation) and FS6.0 (subfield segmentation).

In Sample A, Pearson correlations between raw subregional volumes ranged between \( r = -0.130 \) (fimbria vs. fissure, \( p = 0.756 \)) and \( r = 0.988 \) (CA4 vs. GC-ML-DG, \( p < 1e-20 \)) (median 0.531). Correcting for ICV, age, age-squared, sex, and site led to a slight left-shift (\( r \)-values lowered by \(-0.113 \) on average, resulting median 0.379) within a similar total range \( r = [-0.112; 0.982] \). Correcting for total HV instead of ICV led to a marked left-shift with \( r \) between –0.661 (molecular layer vs. hippocampal tail) and 0.942 (CA4 vs. GC-ML-DG) (median –0.108). This “global correction” effect is best explained by the strong collinearity between subfield volumes and the total HV, even after correction for ICV, age, age-squared, sex, and site (\( r = [0.327; 0.974] \)) (see Supplemental Figure 1a).

In Sample B, correlations between raw volumes were similar (\( r \) between 0.095 [fimbria vs. hippocampal tail, \( p = 0.70 \)] and 0.994 [CA4 vs. GC-ML-DG, \( p < 1e-20 \)], median 0.575). Partial correlation coefficients between subregional volumes corrected for the same covariates as in Sample A were again mildly left-shifted (range \( r = [-0.085; 0.992] \), median 0.475). For comparability with published cross-correlations of residualized subregional volumes (van der Meer et al., 2018), the age-squared term was omitted which changed the distribution only marginally.
### TABLE 2  Overview of reported FS-based hippocampal subfield studies, 01/2013-12/2019. Studies from the first category (heritability studies) are summarized in Section 4.2. The clinical/behavioral/biomarker field was categorized into 15 domains, and the genetic field into 4 domains. Minimum keywords are given for single studies or groups of studies to characterize subdomains

<table>
<thead>
<tr>
<th>Domain</th>
<th>Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Heritability</strong></td>
<td>Heritability (Greenspan, Arakelian, &amp; van Erp, 2016; van der Meer et al., 2018; Whelan et al., 2016), heritability and genetic correlation (Elman et al., 2019)</td>
</tr>
<tr>
<td><strong>Clinical diagnoses, behavioral and cognitive phenotypes, or other biomarkers</strong></td>
<td></td>
</tr>
<tr>
<td>Neurodevelopmental studies</td>
<td>Cognitive aspects during early childhood until early adulthood (Krogsrud et al., 2014; Riggins et al., 2018; Tamnes et al., 2014)</td>
</tr>
<tr>
<td>Healthy aging</td>
<td>Cognitive performance (Aslaksen, Bystad, Ørbo, &amp; Vanberg, 2018; Delazer et al., 2019; Engvig et al., 2012; Ezzati, Katz, Lipton, Lipton, &amp; Verghese, 2015; Foster, Kennedy, Hoagye, &amp; Rodrigo, 2019; Pereira et al., 2014; Zammit et al., 2017; Zheng, Cui, et al., 2018; Zheng, Liu, et al., 2018), hormonal influences (Pintzka &amp; Häberg, 2015), protective effects of education (Jiang, Cao, et al., 2019)</td>
</tr>
<tr>
<td>Other intermediate aging phenotypes: Beta-amyloid and tau pathology (Caldwell et al., 2018; Hsu et al., 2015; Lindberg et al., 2017; Parker, Cash, et al., 2019; Parker, Slattery, et al., 2019; Rahayel et al., 2019)</td>
<td></td>
</tr>
<tr>
<td>Memory functions in homeless and marginally housed persons (Gicas et al., 2019)</td>
<td></td>
</tr>
<tr>
<td>Subfields as ROI for fMRI (memory formation) (Thavabalasingam, O’Neil, Tay, Nestor, &amp; Lee, 2019)</td>
<td></td>
</tr>
<tr>
<td>Pathological aging</td>
<td>Amnestic or vascular MCI, established AD or LBD (de Flores et al., 2015; Evans et al., 2018; Gomar et al., 2017; Györfi et al., 2017; Hirjak, Wolf, et al., 2017; Hirjak, Sambataro, et al., 2019; Kälin et al., 2017; Kang, Lim, Joo, Lee, &amp; Lee, 2018; Khan et al., 2015; Li et al., 2016; Li, Dong, Xie, &amp; Zhang, 2013; Liang et al., 2018; Lim et al., 2012; Lim et al., 2013; Mak et al., 2016; Mak et al., 2017; Parker, Slattery, et al., 2019; Sarica et al., 2018; Shim et al., 2017; Wang, Yu, et al., 2018)</td>
</tr>
<tr>
<td>Vitamin D in MCI (Al-Amin et al., 2019), IL-4 levels in MCI and AD (Boccardi et al., 2019)</td>
<td></td>
</tr>
<tr>
<td>Spectral analyses (including healthy elderly, subjective memory complaints, MCI and AD; risk of conversion from MCI to AD) (DeVivo et al., 2019; Grajski &amp; Bressler, 2019; Izzo, Andreassen, Westlye, &amp; van der Meer, 2019; Marizzoni et al., 2019; Zhao et al., 2019)</td>
<td></td>
</tr>
<tr>
<td>Other neurodegenerative conditions</td>
<td>PD (including dementia) (Foo et al., 2017; Lenka et al., 2018; Low, Foo, Yong, Tan, &amp; Kandiah, 2019; Park et al., 2019; Pereira et al., 2013; Stav et al., 2016; Uribe et al., 2018)</td>
</tr>
<tr>
<td>Frontotemporal dementia (Bocchetta et al., 2018), LBD, and beta amyloid (Mak et al., 2019)</td>
<td></td>
</tr>
<tr>
<td>Multisystem atrophy and PD (Wang, Zhang, Yang, Luo, &amp; Fan, 2019)</td>
<td></td>
</tr>
<tr>
<td>Amyotrophic lateral sclerosis (Christidi et al., 2019), primary lateral sclerosis (Finegan et al., 2019)</td>
<td></td>
</tr>
<tr>
<td>Epilepsy</td>
<td>Mesial temporal lobe epilepsy (Costa et al., 2019; Donos et al., 2018; Duarte et al., 2018; Kim, Suh, &amp; Kim, 2015; Kreilkamp, Weber, Elkommos, Richardson, &amp; Keller, 2018; Lee, Seo, &amp; Park, 2019; Long, Feng, Liao, Zhou, &amp; Urbin, 2018; Peixoto-Santos et al., 2018; Schoene-Bake et al., 2014; Sone et al., 2016)</td>
</tr>
<tr>
<td>Subfields as ROI for fiber tracking (Rutland et al., 2018)</td>
<td></td>
</tr>
<tr>
<td>CNS (autoimmune) inflammatory disorder</td>
<td>Multiple sclerosis (González Torre et al., 2017; Zuppicchini &amp; Sandry, 2018), clinically isolated syndrome (Cacciaguerra et al., 2019), neuromyelitis optica spectrum disorder (Chen et al., 2019), anti-LGI1 encephalitis (Finke et al., 2017), anti-NMDA receptor encephalitis (Finke et al., 2016), subtypes of limbic encephalitis (Ernst et al., 2019)</td>
</tr>
<tr>
<td>Neurotoxicity</td>
<td>Toxic agents, radiotherapy or hypoxia (Decker et al., 2017; Lv et al., 2018; Ørbo, Vanberg, Tande, Anke, &amp; Aslaksen, 2018; Phillips et al., 2020; Stamenova et al., 2018), alcoholism (Lee et al., 2016; Mole, Mak, Chien, &amp; Voorn, 2016; Zahr, Pohl, Saranathan, Sullivan, &amp; Pfefferbaum, 2019), alcohol withdrawal (Kühn et al., 2014), familial risk for alcohol use disorder (adolescents) (Maksimovskiy et al., 2019), cannabis use (Beale et al., 2018), cigarette smoking (Durazzo, Meyerhoff, &amp; Nixon, 2013), preterm birth in school-aged children (Aanes et al., 2019)</td>
</tr>
<tr>
<td>Stress response</td>
<td>Plasma markers of oxidative stress (van Velzen et al., 2017), peripheral inflammatory markers in HIV (Fleischmann et al., 2018) expression of glucocorticoid inducible genes in MDD (Frodell, Carballo, et al., 2014)</td>
</tr>
<tr>
<td>Socioeconomic status and chronic physiological stress (hair cortisol) (Merz et al., 2019), Perceived stress (Zimmerman et al., 2016)</td>
<td></td>
</tr>
<tr>
<td>MDD</td>
<td>Female MDD patients (Han, Won, Sim, &amp; Tae, 2016; Kühn et al., 2012), ECT and treatment response to ECT (Cao et al., 2018; Gryglewski et al., 2019) or antidepressants (Hu et al., 2019; Maller et al., 2017)</td>
</tr>
<tr>
<td>Acute and remitted depression (Kraus et al., 2019), depression symptom severity (Brown et al., 2019)</td>
<td></td>
</tr>
<tr>
<td>Neurovascular disease in late onset MDD (Choi et al., 2017), MDD and aging (Szymkowicz et al., 2017), MDD and interleukin-6 (Kakeda et al., 2018), MDD and other inflammatory markers (Lindqvist et al., 2014)</td>
<td></td>
</tr>
<tr>
<td>Subfields as ROI for fiber tracking (Doojin et al., 2018)</td>
<td></td>
</tr>
<tr>
<td><strong>Subfields as ROI for fMRI (memory formation)</strong> (Thavabalasingam, O’Neil, Tay, Nestor, &amp; Lee, 2019)</td>
<td></td>
</tr>
</tbody>
</table>

**SÄMANN ET AL.**
TABLE 2 (Continued)

<table>
<thead>
<tr>
<th>Domain</th>
<th>Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bipolar disease</td>
<td>Bipolar disease (Elvsåshagen et al., 2013, 2016), comparison with MDD in adults (Cao et al., 2017; Han et al., 2019) and in children and adolescents (Tannous et al., 2018), lithium effects (Giaumomatos et al., 2015; Hartberg et al., 2015; Simonetti et al., 2016). Predominant polarity (Janiri, et al., 2019).</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>Schizophrenia (Zheng et al., 2019), first episode and chronic disease (Kawano et al., 2015; McHugo et al., 2018), symptom correlates (Han et al., 2016; Kühn et al., 2012), first episode psychosis (Baglio et al., 2018; Buchy et al., 2016; Li et al., 2018), response to electroconvulsive therapy (Jiang, Xu, et al., 2019), conversion of high risk patients (Provenzano et al., 2020), young relatives at risk (Francis et al., 2013; Ho, Iglesias, et al., 2017; Ho, Holt, et al., 2017), progression patterns (Ho, Holt, et al., 2017; Ho, Iglesias, et al., 2017), genetic and cognitive correlates of schizophrenia (Nakahara et al., 2019). Metacognition and insight deficits (Alkan, Davies, Greenwood, &amp; Evans, 2019; Hýža, Kuhn, Česková, Ustohal, &amp; Kaspárek, 2016; Orfei et al., 2017).</td>
</tr>
<tr>
<td>PTSD and early life adversity</td>
<td>PTSD (Averill et al., 2017; Bæn et al., 2014; Hayes et al., 2017; L. Chen et al., 2016), childhood maltreatment (Chalavi et al., 2015; Teicher, Anderson, &amp; Polcari, 2012), verbal abuse (Lee et al., 2018), childhood trauma in schizophrenia and healthy controls (du Plessis et al., 2019), in bipolar disorders and healthy controls (Janiri, Sani, et al., 2019) and in adolescence (Malhi et al., 2019).</td>
</tr>
<tr>
<td>Other neuropsychiatric conditions</td>
<td>Pain symptoms (Ezzati et al., 2014), ADHD (Al-Amin, Zhinchenko, &amp; Geyer, 2018), anorexia nervosa (Myrvang et al., 2018), obsessive compulsive disorder (Zhang et al., 2019), and transient global amnesia (Wang, Zhang, et al., 2018), essential tremor (Prasad et al., 2019), rapid eye movement sleep behavior disorder (Campabadal et al., 2019), trigeminal neuralgia (Vuculik, Noorani, Hung, &amp; Hodaie, 2019). Thalamic infarction (Chen et al., 2016), sequelae of microsurgical aneurysm clipping (Hedderich et al., 2019).</td>
</tr>
<tr>
<td>Systemic disease</td>
<td>Prediabetes (Dong et al., 2019) primary biliary cholangitis (Mosher et al., 2018), systemic lupus erythematosus (Bödi et al., 2017).</td>
</tr>
<tr>
<td>Transdiagnostic approach</td>
<td>Psychosis spectrum (Francis et al., 2013; Haukvik et al., 2015; Mathew et al., 2014; Vargas et al., 2017), unipolar/bipolar spectrum in adults (Cao et al., 2017; Han et al., 2019) or children and adolescents (Tannous et al., 2018), Bipolar disease and schizophrenia (pooling different subfield methods) (Haukvik, Tamnes, Söderman, &amp; Agartz, 2018). Social anxiety disorder, childhood trauma and PTSD (Ahmed-Leitao et al., 2019).</td>
</tr>
<tr>
<td>Genetic or epigenetic candidate analyses and exploratory imaging genetics analyses</td>
<td>Candidate SNPs/genes: BDNF val66met variants (Aas et al., 2014; Frodl, Skokauskas, et al., 2014), TESC gene and MDD (Han et al., 2017); COMT and first-episode MDD (Otsuka et al., 2019), oxytocin receptor gene and MDD (Na et al., 2018), 22q11.2 deletion syndrome (Mancini et al., 2019).</td>
</tr>
<tr>
<td>Epigenetics and gene–environment effects</td>
<td>Epigenetic modifications of glucocorticoid receptor in MDD and controls (Na et al., 2014), environmental adversity and COMT, BDNF, and 5-HTTLPR (Rabl et al., 2014).</td>
</tr>
<tr>
<td>Polygenic risk</td>
<td>Aerobic exercise and polygenic risk for schizophrenia (Papiol et al., 2017; Papiol et al., 2019).</td>
</tr>
<tr>
<td>GWAS</td>
<td>GWAS of all subfields (van der Meer et al., 2018), GWAS of the dentate gyrus (Nakahara et al., 2019), GWAS with a methodological focus (Couvy-Duchesne et al., 2019).</td>
</tr>
</tbody>
</table>

Abbreviations: AD, Alzheimer’s disease; ECT, electroconvulsive therapy; FS, FreeSurfer; GWAS, Genome wide association analysis; LBD, Lewy body dementia; MCI, mild cognitive impairment; MDD, major depressive disorder; PD, Parkinson’s disease; PTSD, posttraumatic stress disorder; ROI, region of interest.

$r = [-0.074; 0.992], median 0.471$ and basically matched the mentioned report. Correcting for total HV again led to a strong left shift of the distribution $r = [-0.663; 0.964], median -0.024$ (see Supplemental Figure 1b).

### 2.3.1 Comparing the subregional volumes versus total HV correlations (referred to as “autocorrelations”) between Samples A and B

For this analysis, Sample A was restricted to control subjects only for better comparability ($N = 209$, age range 19–79 years). Similar autocorrelation ranges were detected in both samples after correcting for age, age-squared, sex, ICV, and site (Sample A: $[0.201; 0.967]$ vs. Sample B: $[0.390; 0.979]$). After Fisher’s $z$ transformation, no difference was detected between the samples (two-sided two-sample $t$ test, unequal variances assumed, $p = .385$). Six of twelve subregions (three with the lowest, and three with the highest correlations) were equally ranked between the samples, whereas for the remaining subregions the ranking positions were different (Table 1). Still, the correlation between the corresponding Fisher’s $z$ scores of the two samples was high ($r = 0.967$).

### 2.3.2 Autocorrelations and subfield size

The degree of autocorrelation of a subfield was linearly correlated with the average raw volume ($r = 0.766$ [Sample A], and $r = 0.690$ [Sample B]).
2.3.3 | Interpretation

The concordance between the autocorrelation of subfields with their size likely reflects the above (Section 2.2.3) summarized lower measurement reliability of smaller subfields. Although imperfect in design, as measurements were not repeated in the same subjects, these cross-sectional results point out that subregion volumes at 1.5 Tesla do not show a systematically stronger influence of the probability atlas over the contrast in the input data compared to 3 Tesla. Still, ranking shifts were observed for half of the subfields between platforms, highlighting that pooling results from different field strengths (or highly different acquisition schemes) is not recommended; rather, these should be treated as distinct samples.

2.3.4 | Implications for multiple test correction

For multiple testing corrections, given the demonstrated and reported high correlation among subfield volumes (Elman et al., 2019; van der Meer et al., 2018), Bonferroni correction might be too strict. More recently, it was estimated that the 12 subfield volumes, after correction for covariates including ICV, represent seven effectively independent tests (software available at https://neurogenetics.qimrberghofer.edu.au/matSpDLite) (Couvy-Duchesne et al., 2019; Li & Ji, 2005; Li, Yeung, Cherry, & Sham, 2012). False discovery rate based correction methods as used in several studies so far seem a good option. More complex methods that take into account the actual correlation pattern like approximations to permutation analysis (Conneely & Boehnke, 2007; Zugman et al., 2020) or spectral decomposition (Nyholt, 2004) can be considered.

3 | QC ASPECTS

3.1 | Reported QC procedures

QC procedures were mentioned in about 44% of 162 reviewed reports (Supplemental Table 1, see Section 4 for details on the review procedure), with heterogeneous elaborateness and sometimes involving up to three independent raters. This does not necessarily indicate that no QC was performed in the remaining studies, as for most imaging laboratories, plausibility checks are expected and may thus be underreported. The overall technical success rate of the subfield segmentation is influenced by installation preconditions and rarely explicitly stated (e.g., 96.5% for FS6.0 in Schmidt et al. (2018)), so cannot be reliably quantified due to the heterogeneity of reports but only estimated to be over 95%. Most QC descriptions point to a visual and mostly qualitative, holistic assessment. Ten studies (~6%) excluded cases only based on outlier features (e.g., > 5 SDs), the largest being the study of van der Meer et al. (2018) (see Supplemental Table 1, column QC details). In turn, most studies used outlier features to guide visual inspection but did not exclude outliers without identifying a technical segmentation failure. Five studies (~3%) report manual editing with no further details (see Supplemental Table 1, column QC details). Only exceptionally, cases were excluded due to complete failure of the segmentation, some of them related to gross brain pathology. Several studies mentioned FS viewing tools to inspect segmentations, or mention the ENIGMA protocol, or leave this unspecified. There appears to be a consensus that in the case of peculiarities or segmentation failure, the whole case is excluded (at least the affected hemisphere), and not single subfields.

3.2 | Theoretical considerations for QC

The FS6.0-based hippocampal segmentation algorithm is fully automated and manual corrections are technically feasible, but not designated. Still, the algorithm should not be used in a completely “unsupervised way,” as segmentation failures in the complex, multistep procedure may occur for several reasons. The FS algorithm as such is very robust, and depending on basic checks before starting the segmentation, a map with hippocampal subfield volumes is generally obtained. The FS tool package allows one to overlay the hippocampal segmentation output on a background image with a simple command, usually in native space, for visual QC (Figure 2b), which is useful in smaller studies.

Still, there is no standardized QC procedure for the hippocampal segmentation tool in FS. While requiring a certain amount of time, a visual QC step combined with an automated recommendation of critical cases, appears to be a good compromise to control the risk of technical or algorithmic problems that—if undetected—may lead to distorted results and conclusions. Based on the Bayesian approach of the probabilistic atlas mesh being matched with the actual data, the output represents a data-informed modification of this mesh, similar to other deformation-based procedures, as in tensor-based morphometry, for example. As they are explicitly penalized, extreme deviations of one subregion (in terms of 3D shape/surface and eventually volume) are thus unlikely, and will not occur independently of other subregions. The QC rater, in our opinion, needs no formal neuroradiological training, yet should be sufficiently informed to distinguish between:

1. normal anatomical variants that occur in the hippocampal area and might adversely affect the automated algorithm;
2. minor, visually conspicuous, yet volumetrically negligible phenomena;
3. incomplete segmentations or otherwise atypical segmentation results.

Category (1) includes, for example, cystic formations with CSF intensity within or close to the hippocampal formation, such as sulcus remnant cysts (also referred to as the “hippocampal sulcus residual cavity”) (Li et al., 2006; Sasaki, Sone, Ebara, & Tamakawa, 1993) or choroidal fissure cysts (that can, histologically, represent arachnoid, neuroglial, or neuroepithelial cysts) (Osborn & Preece, 2006). These might stand out on the raw image, the fissure overlay or the full segmentation overlay, yet with little effect on the validity of the actual adjacent subfield volume measures. Following this logic, most subjects
FIGURE 3 Examples of normal anatomical variants, peculiarities and phenomena, and discrepancy between the binary hippocampal mask and segmentation results. Images are displayed at (unsmoothed) 1-mm isometric resolution, and subfield overlays at 0.333-mm isometric resolution. Anonymized examples are differently contrasted as they stem from several sites with different display settings. All shown examples would not necessarily need to be excluded, yet see specific comments on a2 and a5. (a1) Example of sulcus remnant cysts that are found in ~25% of adults between the dentate gyrus and the cornu ammonis; considered an incidental finding with no pathological implication; classified as hippocampal fissure (yellow on mid row image) which represents CSF intensity. Similar peri- or intrahippocampal cysts exist, such as choroidal fissure cysts (etiologically arachnoid or neurogial or neuroepithelial cysts) may similarly be classified as hippocampal fissure, or extra-hippocampal (“background”) by the algorithm, depending on location details. Certain types of cysts may co-occur with enlarged perivascular spaces (Virchow Robin spaces), for example, in the lower basal ganglia or midbrain (see open arrows in a1). (a2) A cystic area not classified as fissure, but as extra-hippocampal background, and a part of CA1 seems neglected. Depending on the amount of such truncated hippocampal tissue, extreme cases of this type should be excluded. (a3–a5) Examples of cysts either classified clearly as fissure, or as background (i.e., no hippocampal subregion). a5 likely classified the fissure correctly that, however, appears brighter than CSF on the raw image due to partial volume effects. More extreme cases (hippocampal fissure intensity being too bright on T1WI or not represented on T2WI) should be excluded, and the hippocampal fissure volume may contain cysts. (b) Very frequent (90%) yet practically negligible observed discontinuous appearance of CA1 and/or smaller spared regions (“holes”) within CA1. (c) Subfields extending into another subfield or forming small extensions/islands; though visually conspicuous the volume effect of such extensions is negligible. (d) “Holes” localizable to fimbria (light pink, d1), or CA3 (green) or CA1 (red) as in d2 or d3 are caused by the removal of the alveus (white matter layer on the superior rim of the hippocampus) area in the final binary classification step, which may cause the impression of an abrupt ending of the segmentation. (e) “Bulky” appearance of CA1 in the sense of this subfield strongly dominating the appearance on one slice. This may occur due to strict orthogonal slicing in native space. (f1) Example of standard hippocampal segmentation (recon-all) missing parts of the posterior hippocampus, while the subfield segmentation v6.0 correctly detects these parts. Incompleteness of the standard binary hippocampal starting mask should therefore not be an exclusion criterion per se. (f2) Anterior parts seem truncated, yet lie in the amygdala complex; this appearance is normal and does not indicate a failed segmentation

in this category do not need to be excluded. Research on the effect of known normal variants is ongoing; more recently, a correlation between CA1 volumes and incomplete hippocampal inversion (IHI), a very common normal variant found in about 20% of healthy subjects (Colenutt, McCann, Knight, Coulthard, & Kauppinen, 2018), has been reported
Category (2) represents phenomena that could be referred to as “small artifacts” with no critical effect on the resulting volumes. Yet, it seems useful to identify examples in this category, particularly for raters assessing a smaller study that may only contain very few examples of such phenomena (see Figure 3b–e and legends for specific examples).

Category (3) is potentially complex and comprises heterogeneous underlying causes, including technical failures (e.g., algorithm failure, or incorrect orientation of the input image), low input image quality, or extreme pathology with gross atrophy or intrahippocampal signal intensity changes. Particularly the latter category—extreme pathology—may lead to a decision to keep the subject in the sample, or exceptionally exclude only selected subfields. Such a scenario may for example be marked postoperative or posttraumatic changes of the medial temporal lobe that prevent proper detection of certain subfields, strong intrahippocampal intensity changes caused by hippocampal sclerosis, ischemia or encephalitis, or strongly progressed AD or other forms of neurodegeneration. For this reason, the underlying T1WI is displayed in parallel.

Figure 3 covers examples of Categories (1) and (2). Additional examples of deviations may be found in Figure 12 and supplemental Figure 22 of the original report on the ex vivo atlas-based method (Iglesias et al., 2015).

Two further points deserve attention: First, circularity is problematic—in the sense of (visually) rating those features that will be used as the dependent variable of the analysis. In addition, as the hippocampal segmentation is based on—though not fully dependent on—a proper general segmentation of the hippocampus (as part of the recon-all command in FSL), such a control step should be incorporated in the QC (see below). For our QC pipeline, we thus developed a triple overlay of (a) the resulting hippocampal fissure (an output of the subfield segmentation) on (b) the starting hippocampus mask of the general segmentation, both overlaid on (c) the T1-weighted background image as the basis for the visual QC. In addition to this, yet not in isolation, the full set of hippocampal subregions is shown (Figure 2). Second, the general cortical and subcortical segmentation should also undergo a QC, for example, by the packages supported online, as of April 2017 (http://enigma.ini.usc.edu/protocols/imaging-protocols), for two main reasons: (a) the generation of intensity distributions is based on defined white matter areas and successful cortical/subcortical segmentation and (b) global measures such as ICV, total brain volume or total gray matter (TGM) are usually needed for the statistical model and rely on a proper general registration or segmentation.

As both the general QC and the subfield QC show example slices of the original input image, this should help in detecting very low-quality input images (e.g., blurred margins due to motion artifacts). Such images increase the risk of general segmentation failure (e.g., parts of lobes missing, or considerable underestimates of cortical thickness) and/or incomplete progression of the hippocampal segmentation to the borders of the hippocampus. The latter may be found in case of an extremely degenerated hippocampus where internal CSF signals may mimic a true margin. In our experience, such global low-quality cases often lead to strongly underestimated TGM values or TGM/ICV ratios that appear as outlier values, particularly when plotted against age.

3.3 | A proposed QC pipeline

Within the ENIGMA consortium, we have developed a standardized set of MATLAB, Linux shell, and R scripts for QC of hippocampal subfields derived from the F56.0 algorithm that provide several functions, including (a) starting subfield segmentations on a group of subjects that have undergone the general recon-all segmentation, (b) gaining a standardized readout table of subregional volumes and global volumes from the general segmentation, (c) a list of cases that should be inspected individually, and (d) a browsable hypertext markup language (HTML) file that allows to perform this visual QC also in larger samples (Figure 4).

Steps (a) and (b) are based on Linux shell scripts. Step (c) is an R script based on two principles: First, an outlier calculation scheme (+/- 2.98 SDs, each representing the roughly 1% of cases in the tails of a normal distribution) directed to all subfield volumes, yet also total brain volume, total GM volume, ICV, and GM/ICV ratio. Second, as the rank order of the subregional volumes is very robust (Figure 2a) across MR platforms and samples, a set of three violations of this ranking order has been defined from an analysis of N = 626 subjects in which the deviation pattern of the actual rank from the group mean rank was calculated: (a) CA1 not being the largest subregion (i.e., not rank #1; frequency ~ 11%), (b) hippocampal tail being below rank #3 (frequency ~2%), and (c) subiculum not being ranked #4 (frequency ~ 2%). The first rule is optional (switched off by default) as it may be oversensitive and too unspecific in larger samples.

For Step (d), the number of cases summarized per HTML-file can be defined by the user. In this way, several separate files are generated, for example, for distribution to several raters. Further, a random order mode (to avoid QC order biases) can be requested, or a sparse version of the HTML output with fewer example slices that allows one to accelerate the throughput in very large samples (e.g., N > 1,000).

Overall, when using our ENIGMA Hippocampal Subfields QC protocol, we recommend the following steps, as also shown in Figure 5, before statistical analyses of subregional volumes:

1. Perform general FS segmentation and append the hippocampal segmentation module, ideally using the same version (currently 6.0 or 7.0) (Note: Combining FS5.3 general segmentation with FS6.0 subfield version is feasible and produces highly similar results, compared to FS6.0 only; also see supplemental material in van der Meer et al. (2018).)

2. Perform a QC of the general cortical and subcortical segmentation, using the respective ENIGMA scripts and instructions (http://enigma.ini.usc.edu/protocols/imaging-protocols).

3. Read out hippocampal subregional volume and general segmentation results using the script from the SUBFIELDS package: run the R script outliers_hippo_and_QC_support.R on this output to generate a list of cases that could harbor problems (see Figure 5 for an example).
4. Run the three QC preparation scripts to eventually generate the QC-HTML files; adapt graphical and random order or sparse output settings in the scripts as required.

5. Inspect all datasets in the QC-HTML files with a standard browser, with particular attention to the flagged datasets. Additional recommendations are:
   a. To become familiar with the study specific variance of segmentations and to avoid an order bias, it may be efficient to view a random selection of cases at a higher pace, collecting “global impressions” first.
   b. Before browsing the static HTML-file and if unfamiliar with the FS hippocampal segmentation output, it may be beneficial to explore a handful of cases using the FS viewer Freeview that allows scrolling in all three dimensions. As FS output is in native space, this step is useful to become acquainted with the variance caused by imperfectly corresponding anatomical slice positions between subjects.

4 | FS-BASED HIPPOCAMPAL SUBFIELD ANALYSIS: STATUS OF CLINICAL NEUROSCIENCE PUBLICATIONS

The studies referred to in this section were retrieved from a PubMed literature search performed on December 14, 2019. The search combination was “hippocampal subfields” OR “hippocampal subregions” OR “hippocampal subfield” OR “hippocampal subregion” AND “MRI”
Ideally first motion artefact/general quality QC (e. g. on the PACS)

1. Run general segmentation (recon-all)
   
   QC of general cortical/subcortical segmentation:
   [http://enigma.ini.usc.edu/protocols/imaging-protocols](http://enigma.ini.usc.edu/protocols/imaging-protocols)
   + Identify null results, rotation/image format errors & re-run

2. Run FreeSurfer hippocampal subfield segmentation (with respective flags)

3. Extract subfields into table and look for null outputs
   
   Null outputs: check input *mgz image (freeview), screen for Matlab breakdown based failures or cluster problems & re-run subfield algorithm

5. Run R script outliers_hippo_and_QC_support.R

   Review histograms including ICV:
   Strongly skewed? Extreme outliers?
   Plot age vs. total GM/ICV

6. Run preparation scripts for visual QC

7. Visual QC of subfield output (based on HTML files gained in step 6)

   **WARMUP** Check 10 cases dynamically (freeview)
   Scroll over samples fast to become acquainted with variance

   **ACTUAL-A** HTML: First global impression approach:
   1. Binary hippocampal mask visible? If not: check recon-all result
   2. Fissure within the mask? If not: freeview + subfields

   **ACTUAL-B** 1. Strong truncations comp’d with mask? freeview + subfields
   2. Subfield related peculiarities:
      - Cystic lesions: - Classified as fissure or background? KEEP
      - Discontinuities: - Explained by bordering alveus? KEEP
      - Not alveus related: freeview + subfields

( limited [filter] to human studies and a time frame January 1, 2013–December 14, 2019), resulting in 401 reports. All 401 abstracts were reviewed (P. G. S.) and those in which the segmentation technique was not the automated FS tool (i.e., other automated tools or manual segmentation protocols) were excluded. Reviews and retrospective meta-analyses were excluded to avoid double inclusion of original studies. After excluding reliability and purely methodological studies that we refer to in Section 4.2, 162 reports were available for further analyses.

From these studies, we aimed to identify domains, topic, and selected methodological information: field strength, MRI sequence (T1WI, T2WI, or both), FS version, global volume correction (if used, and the specific volume used), statistical approach for global correction (regression method, proportion method, covariate method), presence of age and sex as covariates, modeling of nonlinear age effects, and additional covariates. We extracted if a QC of the subfield segmentation was reported, if outliers were excluded and if manual corrections were performed. Finally, we noted if and how hippocampal subfields were (re-)aggregated to create composite markers. Three studies used hippocampal subfield segmentation merely as a tool to define ROIs, leaving 159 studies for the analysis of statistical aspects.

As one goal of ENIGMA is to advance the endophenotype concept (Gottesman & Gould, 2003) for genetic association studies, the returned hippocampal subfield studies were organized into studies of (a) heritability; (b) clinical diagnoses, behavioral and cognitive phenotypes, or other biomarkers; and (c) studies of candidate genes or

---

**FIGURE 5** Overall quality control (QC) flow scheme for a FreeSurfer-based hippocampal subfield study. Depending on local pipelines, general QC steps regarding the raw data quality, motion artifacts and complete coverage may be performed directly on the picture archiving system and be study-independent. A general inspection of the cortical/subcortical segmentation result of FS is recommended before subfield specific operations. Steps 2–7 are supported by a script package written in MATLAB, R, and Shell. “freeview + subfields” refers to an individual check using the interactive FS viewing tool ([https://surfer.nmr.mgh.harvard.edu/fswiki/HippocampalSubfields](https://surfer.nmr.mgh.harvard.edu/fswiki/HippocampalSubfields))
exploratory imaging genetics analyses. Results of (a) are summarized as text in Section 4.1. Results of all Categories (a) and (c) are presented at three different levels of granularity: first, in an iterative process based on the full papers, studies were aggregated into 19 domains (15 clinical/behavioral domains, 4 (epi-)genetic domains) that are depicted in a pie chart (Figure 1, see legend for details on category fusions). Second, in Table 2, subcharacterizations and topics have been added to single studies or groups of studies for a quick overview and references. Third, Supplemental Table 1 lists all 162 studies in more detail focusing on the relevant technical features extracted for the methodological and QC review.

### 4.1 Heritability analyses

Hippocampal subfield volumes from FS6.0 have been analyzed with regard to their heritability—mainly to corroborate their utility as endophenotype—based on monozygotic/dizygotic twin studies (Elman et al., 2019; Greenspan et al., 2016; Whelan et al., 2016) or on genotype information (van der Meer et al., 2018).

Whelan et al. (2016) reported heritability between 56 and 88% for hippocampal subvolumes and total HV, after controlling for age, sex, and age-by-sex effects. Heritability estimates $h^2$ as measured by the proportion of variance in volume attributable to genetics via twin studies were larger than 70% for regions with relatively larger volumes (whole hippocampus, molecular layer, CA1, CA3, CA4, hippocampal tail, granule cell layer, subiculum, and presubiculum) and moderate to high (55% < $h^2$ < 70%) for smaller subregions (HATA, fimbria, parasubiculum, hippocampal fissure). Projection to a hippocampal surface model demonstrated that heritability is larger for posterior subregions (including the hippocampal tail), and smaller for anteroomedial subregions (parasubiculum, presubiculum, fimbria), suggesting a gradient of genetic influences. Even so, the relationship between subregional volume and measurement reliability may also have influenced estimates of heritability. Elman et al. (2019) reported heritability values between 37% (HATA) and 89% (molecular layer) from values controlled for age and sex. Genetic correlations as a measure of shared genetic influences were high between subfields and the total HV, indicating that for standard resolution T1WI, no substantial additional information on the genetic underpinning is gained through the subfield volumes (Elman et al., 2019). Despite finding significant genetic covariance of the subfields, no stable latent genetic traits (composed from combinations of subfields) were found in a factor analysis (Elman et al., 2019).

Greenspan et al. (2016) reported heritability estimates for FS6.0 subfield volumes between 20 and 87% along with high shared genetic variance with total HV (mean 0.79, range 0.50–0.98). After regressing out total HV—or when percentages based on the ratio of the subregional volume/total HV) were considered—heritabilities ranged between 4 and 86% (7 and 84%, respectively). Here, heritability values were not significant for the fimbria, hippocampal fissure and HATA.

From a recent GWAS on hippocampal subregional volumes residualized for ICV, age, sex and site (van der Meer et al., 2018), significant SNP-based heritability was reported from a very large sample (N > 20,000), ranging between 14% (for the parasubiculum) and 27% (for the hippocampal tail).

### 4.2 Application to clinical diagnoses, behavioral, and cognitive phenotypes, (epi-)genetic markers or other biomarkers

We found that hippocampal subfield analyses using the FS tool so far have been performed in a broad range of domains that are quantified in Figure 1 and grouped into subdomains/topics in Table 2. Beyond activity in expected major fields such as hippocampus-dependent cognitive functions, healthy and pathological aging, major psychiatric diseases of the affective/psychosis spectrum and different neurological disorders, a growing interest in psychotraumatology, and neurotoxic influences including systemic inflammation is noted. Subfield volumes are now also probed as endophenotypes, in studies analyzing candidate genes, epigenetic methylation patterns polygenic risk scores, but also exploring the whole genome for associations with single subfields (Nakahara et al., 2019) or all subfields (van der Meer et al., 2018), or for methodological comparisons (Couvy-Duchesne et al., 2019).

### 4.3 Statistical aspects

#### 4.3.1 Modeling global volume effects

One important source of interindividual variance in regional brain volumes is the variation in overall head size (Mathalon, Sullivan, Rawles, & Pfefferbaum, 1993). To increase the specificity of regional results, ICV is often included in the statistical model—of the 159 reports in which subfield volumes were the outcome variable, 78% reported results corrected for ICV. Technically, in FS, ICV is estimated from the matrix created during a linear registration of the individual MRI scan to Talairach space, and is thus also referred to as “estimated total intracranial volume” (Buckner et al., 2004). FS-based studies thus tend to employ this measure as an ICV surrogate. Alternatively, after segmentation (e.g., using Statistical Parametric Mapping [SPM] software), TGM, white matter and CSF volumes can be summed up. Both methods correlate well with the gold standard of manual delineation, yet overestimations have been reported for both SPM8 and FS (version 5.1.0) (Nordenskjöld et al., 2013). Excellent agreement with the gold standard, superior to FS5.3, was recently achieved with SPM12 based segmentation augmented by a manually edited intracranial mask (in MNI space) (Malone et al., 2015). A smaller proportion of studies considered total brain volume (4.4%), supratentorial gray matter volume (2.5%) or total HV (5.7%) as a correction variable. Choosing these variables gradually increases the regional specificity which is particularly relevant if effects in extra-hippocampal brain areas are expected (e.g., neurodegenerative disorders). Correction for total HV allows to study subregional effects that exceed total HV effects, maximizing the regional specificity.
ICV (or another global volume) can be modeled in various ways (O’Brien et al., 2011): first, it can be included as a nuisance covariate in analyses of covariance, linear mixed models or regression analyses, along with other confounding variables (analysis of covariance approach). This approach was used in most (79%) of the 136 studies that included a global volume. Second, volumes can be adjusted by regression analysis (residual approach), so \( \text{vol}_{\text{corr}} = \text{vol} - b \times (\text{ICV} - \text{ICV}_{\text{mean}}) \), where \( b \) is the slope of the regression of vol on ICV (Hayes et al., 2017; Mathalon et al., 1993). About 13% of 136 studies applied this approach. Third, ICV may be used for a direct individual standardization of regional volumes. This approach, applied in 25% of the reports, is referred to as the proportion (or ratio) approach. In its simplest form, \( \text{vol}_{\text{corr}} = \text{vol}/\text{ICV} \). In an extension of this, the sample mean ICV can be multiplied with this term. The advantages and disadvantages of each of these methods are a long discussed topic (Arndt, Cohen, Alliger, Swayze, 2nd, & Andreasen, 1991; Buckner et al., 2004; Jack Jr et al., 1989; O’Brien et al., 2011; Sanfilipo, Benedict, Zivadinov, & Bakshi, 2004; Voevodskaya et al., 2014). No generally valid recommendation can be given that applies to all study types and questions and each approach requires certain assumptions (O’Brien et al., 2011). For example, the covariate allows for a clear attribution of different sources of variance; the regression approach may be unreliable in small samples, but yields a more Gaussian distribution compared to the proportion approach.

An increasingly recognized problem is the possibility of nonlinear (allometric) rather than isometric relationships between total and local cerebral measures (Jäncke, Liem, & Merillat, 2019; O’Brien et al., 2011; Reardon et al., 2016). Toro et al. (2009) demonstrated that assumptions on brain allometry made a critical difference in a genetic study on HV. More recently, van Eijk et al. (2020) reported that allometric scaling may also be found for hippocampal subfield volumes.

### 4.3.2 | Modeling of age effects

Modeling of age effects is another critical point for studies on hippocampal morphology. For TGM volume, when corrected for ICV, the gain in explained variance through nonlinear terms is rather weak (Figure 2d), whereas hippocampal subregion volumes (and other subcortical structures) show nonlinear effects, with a plateau phase in mid adulthood and a decrease starting around age 65 (de Flores et al., 2015) (also see Figure 2c). Similarly, nonlinear developmental trajectories have been reported for the age range between 8 and 30 years for subcortical brain structures including the total hippocampus (Ostby et al., 2009), and for hippocampal subregions between age 4 and 22 years (Krogsgard et al., 2014). This implies that in hippocampal subregional studies of larger age ranges, either guided by theoretical considerations or by data exploration, models with nonlinear terms should be used. In our collection of reports, age effects were modeled in most studies (~89%); four (2.5%) of these studies modeled nonlinear age effects.

In addition, aging processes may be abnormal in clinical conditions, and disease-specific aging processes may falsely translate into group main effects. Respective age-by-sex and age-by-diagnosis interaction terms, and their higher order versions, can help to refine the model and reduce interpretational ambiguity. Here, pooling of multisite data (e.g., mega-analysis) is useful, and methods to remove batch effects under preservation of covariate effects have been suggested for age harmonization (Fortin et al., 2018; Pomponio et al., 2020).

### 4.3.3 | Skipping or reaggregation of hippocampal subfields

Skipping or, in turn, reaggregating hippocampal subfields to create composite measures (differing from those in the original FS6.0 or FS5.3 scheme) was reported in a handful of studies (~4%). Subfields excluded due to their relatively low reliability were mostly the fimbria and the hippocampal fissure (as the latter does not represent hippocampal tissue, but a CSF cleft). Some studies define the subicular complex from the subiculum, the parasubiculum and presubiculum (see Supplemental Table 1, last column, and table legend for details). In FS6.0 and FS7.1, three variants of how the algorithms’ hidden fine-grained subfield parcellation is aggregated to the final labels are provided: (a) a head-body-tail aggregation, useful for comparisons with earlier landmark-based segmentation work or calculation of more coarse ROIs regions, for example, for functional MRI, (b) the aggregation to 12 subregions presented here, and (c) a version called “CA,” in which the internal labels (GC-ML-DG and molecular layer) are absorbed by the CA subfields. More specifically, GC-ML-DG is assigned to CA4, and voxels in the molecular layer are assigned to the closest voxel that is neither in the molecular layer nor background.

McHugo et al. (2018) in their work on psychosis suggested two aggregation schemes for easier comparison with data generated from other commonly used segmentation algorithms (Mueller et al., 2018; Yushkevich et al., 2015) and to account for imperfect boundary definitions for smaller subregions (CA3, GC-DG, molecular layer) in the absence of multispectral input data. These schemes are interesting variants, particularly when volumetric effects along the longitudinal axis of the hippocampus are investigated.

### 4.3.4 | Multivariate analyses and structural covariance analysis

Given the good measurement reliability of most subfield volumes, multivariate analysis including machine learning approaches are particularly attractive, allowing for the detection of complex morphological patterns. For example, hippocampal subfield information predicted treatment outcome in MDD with an accuracy of 83% (Cao et al., 2018). Similarly, linear discriminant analysis using hippocampal subfield information (from both FS5.3 and FS6.0) proved superior to total HV for discriminating AD patients from healthy controls (Iglesias et al., 2015). Both studies add indirect clinical validity to the subfield segmentation.
Structural covariance analysis that is based on the coherence of subfield volumetric patterns across subjects is another interesting approach: For example, Wang, Yu, et al. (2018) and Wang, Zhang, et al. (2018) reported higher covariance between hippocampal subfields and extra-hippocampal gray matter volume in amnestic and vascular MCI patients compared with control subjects. Unsupervised learning tasks such as clustering have revealed groupings within the subfields and with other brain areas (van der Meer et al., 2018). To our knowledge, no clustering of patient samples into subgroups, based on hippocampal subfield patterns, has been reported so far.

5 SUMMARY DISCUSSION AND OUTLOOK

The FS-based hippocampal subfield segmentation algorithm is a widely applied tool that can be used to study physiological hippocampus-dependent processes and pathological conditions. Healthy and pathological aging studies, including the MCI/AD spectrum, together with studies on the psychosis and the affective disorders spectrum make up about 50% of the studies. Early environmental, and genetic influences and studies of neurotoxic influences on hippocampal integrity make up the majority of the remaining studies. The tool provides an opportunity to improve upon the long-known non-specificity of (total) hippocampal volumetry findings in many neuropsychiatric diseases (Geuze, Vermetten, & Brenner, 2005). Due to its easy, automated use, it naturally fosters exploratory studies, and some reports now exist performing GWAS on hippocampal subfield volumes (Morey et al., 2020; Nakahara et al., 2019; van der Meer et al., 2018). Here, recent methodological work has highlighted that for hippocampal subfield volumes multivariate GWAS could be a useful addition to the classical multiple univariate approach (Couvy-Duchesne et al., 2019).

Most clinical/behavioral studies used analysis of covariance or linear mixed models, covarying for age, sex, and ICV. Still, considerable variability exists regarding the modeling of age and global volumes, which hampers retrospective meta-analyses. A similar degree of heterogeneity was noted for QC steps. Prospective harmonized analyses of existing datasets, as performed in ENIGMA, and also the analysis of prospectively acquired subfield data could benefit from our QC protocol as a guideline. Multivariate analyses of hippocampal subfield data are still rarely used but have begun to show promising results, for example, for predicting treatment response in MDD (Cao et al., 2018). In conjunction with good across-site and across-vendor reliability (Brown et al., 2020; Quattrinini et al., 2020), this strengthens the potential of subfield volumes as a biomarker for clinical trials. One possible advantage of subregional volumes over voxel-wise or vertex-wise features is a higher degree of interpretability, as effect weights can be translated back to histology or other levels of validation.

Only very few studies (3.1%) made use of two MRI input channels despite evidence that a second, high resolution T2WI—by detecting critical internal hippocampal boundaries—can improve subfield measurement reliability and validity (Iglesias et al., 2015; Mueller et al., 2018; Winterburn et al., 2013). We suggest that in the currently available FS hippocampal subfield studies the potential of the algorithm is not fully exploited and optimized MRI protocols are considered in future prospective studies.

At least two other automated hippocampal subfield segmentation algorithms exist, both developed and evaluated on 3 Tesla platforms and both based on manual tracings on high resolution MRI images as a reference: MaGET-Brain (multiple automatically generated templates) (Pipitone et al., 2014) and automatic segmentation of hippocampal subfields (ASHS) (Yushkevich et al., 2015). A complex challenge in this methodological field is the definition of an agreed manual segmentation protocol: here, the hippocampal subfields group of the EU Joint Programme—Neurodegenerative Disease Research is aiming at harmonizing a reliable and valid protocol for postmortem data, also fostering its translation to MRI data. A protocol for 7 Tesla MRI data with good agreement between manual and ASHS performance has been published (Wisse et al., 2016), and comparing ex vivo to high-field MRI data (9.4 Tesla) has resulted in further progress in protocol definitions (de Flores et al., 2019). This work is ongoing and will serve as the basis for new validation experiments within the MRI domain (Olsen et al., 2019). The influence of IHI—a common, atypical rotation of the hippocampal formation (Cury et al., 2015) that impacts subfield volumes (Colenutt et al., 2018)—presents another challenge.

As an outlook, hippocampal shape analysis is worthwhile as it can be combined with subfield information to localize shape effects. The shape analysis pipeline of ENIGMA (http://enigma.usc.edu/ongoing/enigma-shape-analysis) is based on the FS subcortical segmentation, followed by topological correction steps and smoothing based on the topology-preserving level set algorithm (Gutman et al., 2015; Xiao Han, Xu, & Prince, 2003). After registration to standard templates (Gutman et al., 2015; Gutman, Wang, Rajagopalan, Toga, & Thompson, 2012), two metrics are calculated for each vertex point: a radial distance that is comparable to a “shape thickness” measure, and the Jacobian determinant (Gutman et al., 2015; Wang et al., 2011), a metric of localized tissue reduction or enlargement of surface area relative to the template shape. Using a subfield surface projection (first presented for F55.3 by Cong et al. (2014)), shape effects can be assigned to specific subfields. Such a combined approach has recently helped to localize hippocampal shape changes in MDD (Ho et al., 2020) and to identify volume changes and a complex bending and displacement of parts of the hippocampus in obsessive–compulsive disorder (Zhang et al., 2019). Beyond this, using a surface-based atlas of hippocampal subvolume projections built on the medial core modeling (Gutman et al., 2012), proportional contributions of volume subfields to a surface effect can be quantified (Figure 6).

In conclusion, the FS-based hippocampal subfield segmentation tool provides a valuable, fine-morphological phenotyping instrument with good measurement reliability and validity. Use of a standardized QC procedure can help to harmonize large-scale, collaborative studies and reduce methodological heterogeneity. Acquisition protocols specifically designed for subfield studies, as well as further external validation and multivariate analysis techniques, could further improve its usefulness in clinical neuroscience.
The modification from FS5.3 to FS6.0 incorporated anatomical and theoretical refinements, as well as a new subregional parcellation that correlates more strongly with gold standards and was derived from higher resolution ex vivo scans. The atlas in FS5.3 had been built from in vivo scans at $0.38 \times 0.38 \times 0.80 \text{ mm}^3$ resolution, which yielded voxels approximately 100 times larger than the ex vivo protocol and required the human labeller to rely on geometric criteria in their delineations. This had led to underestimation of CA1 at the expense of the subiculum and CA2/CA3 (de Flores et al., 2015; Iglesias et al., 2015; Wisse, Biessels, & Geerlings, 2014). Thus, while analyses of between version (FS5.3/FS6.0) agreement have been performed (Whelan et al., 2016), their interpretation is difficult and they are therefore not reported in detail here.

**REFERENCES**


Hu, X., Zhang, L., Hu, X., Lu, L., Tang, S., Li, H., ... Huang, X. (2019). Abnormal hippocampal subfields may be potential predictors of worse early response to antidepressant treatment in drug-Naïve patients with...


Li, W., Li, K., Guan, P., Chen, Y., Xiao, Y., Liu, S., ... Gong, Q. (2018). Volume alteration of hippocampal subfields in first-episode antipsychotic-naïve...


Rabl, U., Meyer, B. M., Diers, K., Bartova, L., Berger, A., Mandler, D., ... Pezawas, L. (2014). Additive gene-environment effects on...


**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section at the end of this article.