



**HAL**  
open science

## How serious is data leakage in deep learning studies on Alzheimer's disease classification?

Junhao Wen, Elina Thibeau-Sutre, Jorge Samper-Gonzalez, Alexandre M Routier, Simona Bottani, Didier Dormont, Stanley Durrleman, Olivier Colliot, Ninon Burgos

### ► To cite this version:

Junhao Wen, Elina Thibeau-Sutre, Jorge Samper-Gonzalez, Alexandre M Routier, Simona Bottani, et al.. How serious is data leakage in deep learning studies on Alzheimer's disease classification?. Organization for Human Brain Mapping (OHBM), Jun 2019, Roma, Italy. hal-03365742

**HAL Id: hal-03365742**

**<https://hal.science/hal-03365742>**

Submitted on 5 Oct 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# How serious is data leakage in deep learning studies on Alzheimer's disease classification?

Junhao Wen<sup>\*1</sup>, Elina Thibeau--Sutre<sup>\*1</sup>, Jorge Samper-González<sup>1</sup>, Alexandre Routier<sup>1</sup>, Simona Bottani<sup>1</sup>, Didier Dormont<sup>1,2</sup>, Stanley Durrleman<sup>1</sup>, Olivier Colliot<sup>1,2,3</sup>, Ninon Burgos<sup>1</sup>

<sup>\*</sup>The authors contributed equally to this study

<sup>1</sup>ARAMIS Lab, ICM, Inserm U1127, CNRS UMR 7225, Sorbonne University, Inria, Paris, France

<sup>2</sup>AP-HP, Department of Neuroradiology, Pitié-Salpêtrière Hospital, Paris, France

<sup>3</sup>AP-HP, Department of Neurology, Pitié-Salpêtrière Hospital, Paris, France

✉ junhao.wen89@gmail.com

✉ elina.thibeausutre@icm-institute.org

@AramisLabParis

In recent years, there has been a strong interest in the use of deep learning (DL) for assisting diagnosis of brain diseases from neuroimaging data. **Unbiased evaluation** of their performances is critical to assess their potential clinical value.

A major source of bias is **data leakage**, that can be difficult to detect for non-specialists.

In this study, focusing on the case of Alzheimer's disease (AD) diagnosis from T1 MRI using convolutional neural networks

(CNN), we performed a rigorous **literature search**, assessed the prevalence of data leakage and analysed its causes. Additionally, we demonstrated the phenomenon of data leakage in a **controlled setting** by focusing on the impact of the data split strategy.

## Methods

### Literature Search

Search engines: **PubMed** & **Scopus**

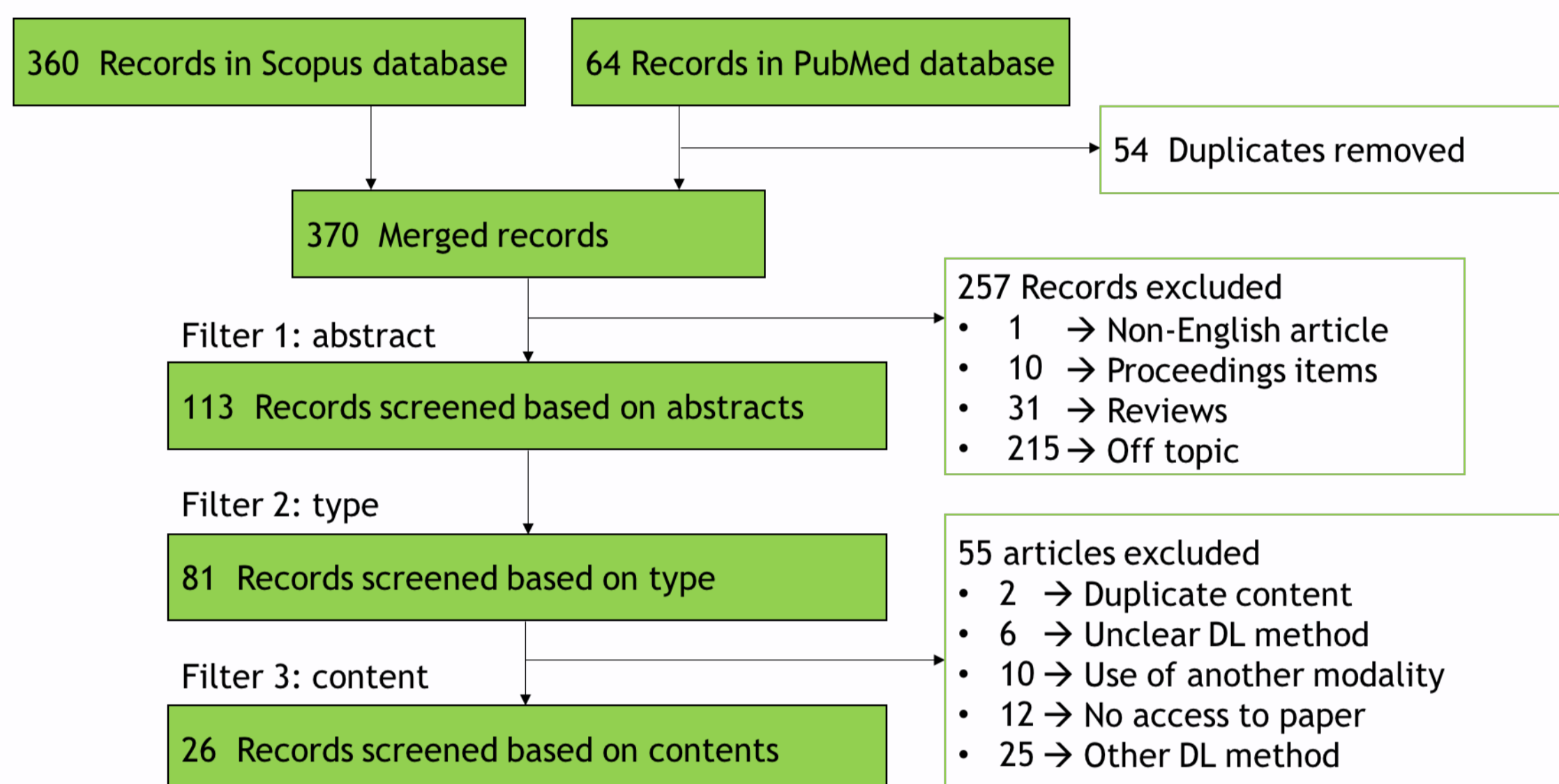


Diagram summarizing the bibliographic methodology.

### Data Leakage

3 categories identified:

- Biased split**  
Data extracted from the same individual is distributed in both the train and the test sets.
- Late split**  
Test / train split is performed after another procedure (feature selection, pretraining...).
- No independent test set**  
The performance is evaluated on the train and / or validation sets.

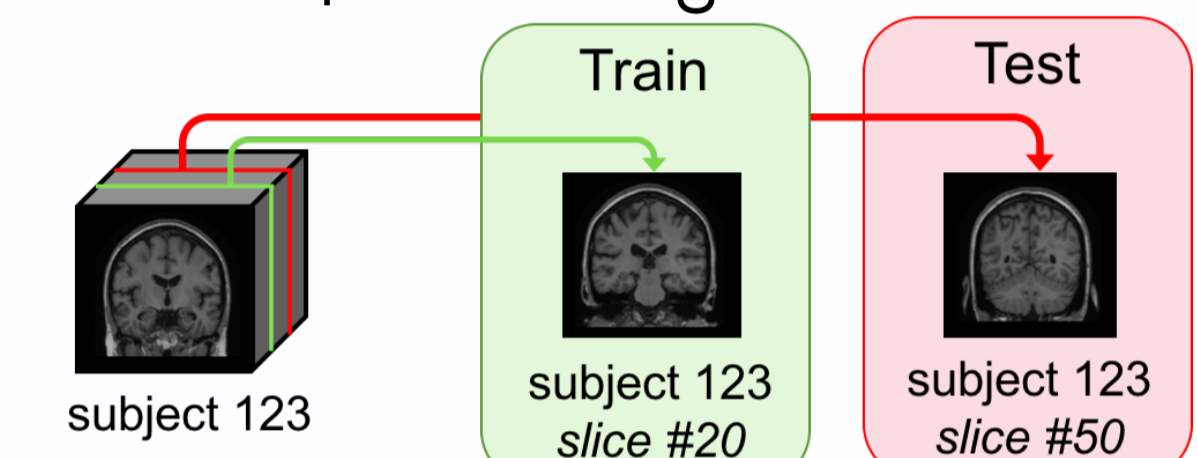
3 possible labels

- Clear** when data leakage is explicitly witnessed
- Unclear** when no sufficient explanation is offered
- None detected** otherwise

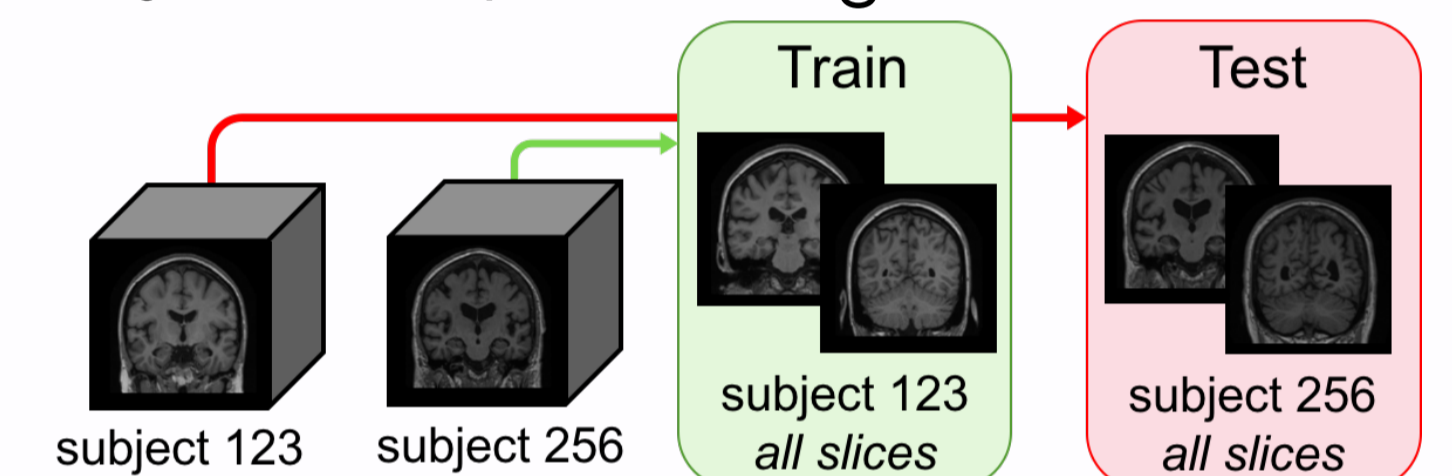
### Application

Evaluation of the impact of biased split: Two experiments were conducted with different data partitioning strategies for the AD vs CN classification on ADNI dataset.

A. Slice-level partitioning



B. Subject-level partitioning



## Results

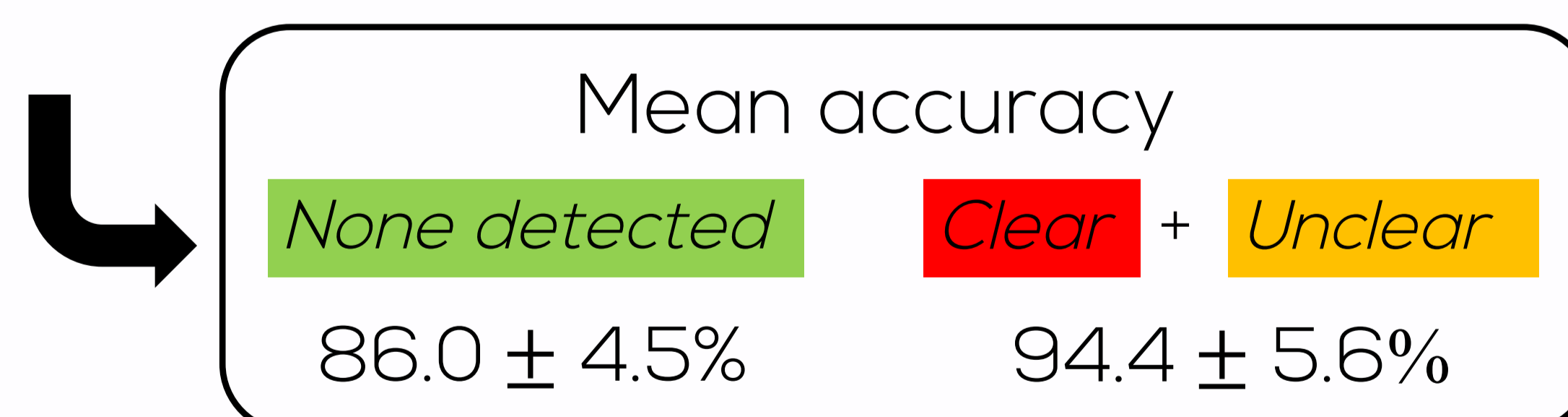
### Summary of the studies performing classification of AD using CNNs on anatomical MRI

A. Studies without data leakage

B. Studies with potential data leakage

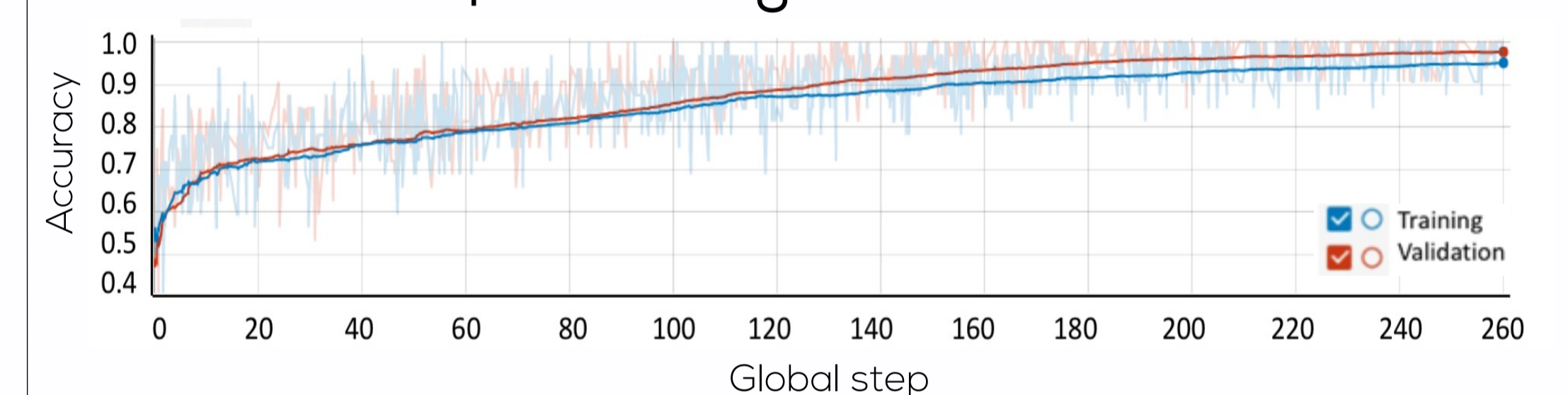
Study	DOI	Accuracy AD vs CN	Data leakage	Study	DOI	Accuracy AD vs CN	Data leakage	Categories		
								1	2	3
Aderghal et al., 2017	10.1007/978-3-319-51811-4_56	83,70%	None detected	Aderghal et al., 2017	10.1145/3095713.3095749	91,41%	Unclear		X	X
Aderghal et al., 2018	10.1109/CBMS.2018.00067	90%	None detected	Hon and Khan, 2017	10.1109/BIBM.2017.8217822	96,25%	Unclear	X	X	
Backstrom et al., 2018*	10.1109/ISBI.2018.8363543	90,11%	None detected	Hosseini-Asl et al., 2018	10.2741/4606	99,30%	Unclear	X		
Cheng et al., 2017	10.1117/12.2281808	87,15%	None detected	Islam and Zhang, 2017	10.1007/978-3-319-70772-3_20	(CN/mild/moderate/severe: 73,75%)	Unclear		X	X
Cheng and Liu, 2017	10.1109/CISP-BMEI.2017.8302281	85,47%	None detected	Taqi et al., 2018	10.1109/MIPR.2018.00032	100%	Unclear			X
Islam and Zhang, 2018**	10.1186/s40708-018-0080-3	(CN/mild/moderate/severe: 93,18%)	None detected	Vu et al., 2017	10.1109/BIGCOMP.2017.7881683	85,24%	Unclear	X		
Korolev et al., 2017	10.1109/ISBI.2017.7950647	80,00%	None detected	Wang et al., 2018	10.1007/s10916-018-0932-7	97,65%	Unclear			X
Li et al., 2018	10.1109/IST.2017.8261566	88,31%	None detected	Backstrom et al., 2018*	10.1109/ISBI.2018.8363543	98,74%	Clear	X		
Li et al., 2018	10.1016/j.compedimag.2018.09.009	89,50%	None detected	Farooq et al., 2017	10.1109/IST.2017.8261460	(AD/LMCI/EMCI/ CN: 98,88%)	Clear	X		
Liu et al., 2018	10.1007/s12021-018-9370-4	84,97%	None detected	Gunawardena et al., 2017	10.1109/MZVIP.2017.8211486	(AD/MCI/CN: 96%)	Clear	X	X	
Liu. et al., 2018	10.1016/j.media.2017.10.005	91,09%	None detected	Vu et al., 2018	10.1007/s00500-018-3421-5	86,25%	Clear	X		X
Liu. et al., 2018	10.1109/JBHI.2018.2791863	90,56%	None detected	Wang S. et al., 2017	10.1007/978-3-319-68600-4_43	(MCI/CN: 90,60%)	Clear	X	X	
Senanayake et al., 2018	10.1109/ISBI.2018.8363832	76%	None detected	When different from AD vs CN, the classification task is specified in brackets.						
Shmulev et al., 2018	10.1007/978-3-030-00689-1_9	(sMCI/pMCI: 62%)	None detected	* (Backstrom et al., 2018) studied the impact of a biased data split						
Valliani and Soni., 2017	10.1145/3107411.3108224	81,30%	None detected	** Use of imbalanced accuracy on an imbalanced dataset						

Data leakage categories:  
 1. Biased split  
 2. Late split  
 3. No independent test set

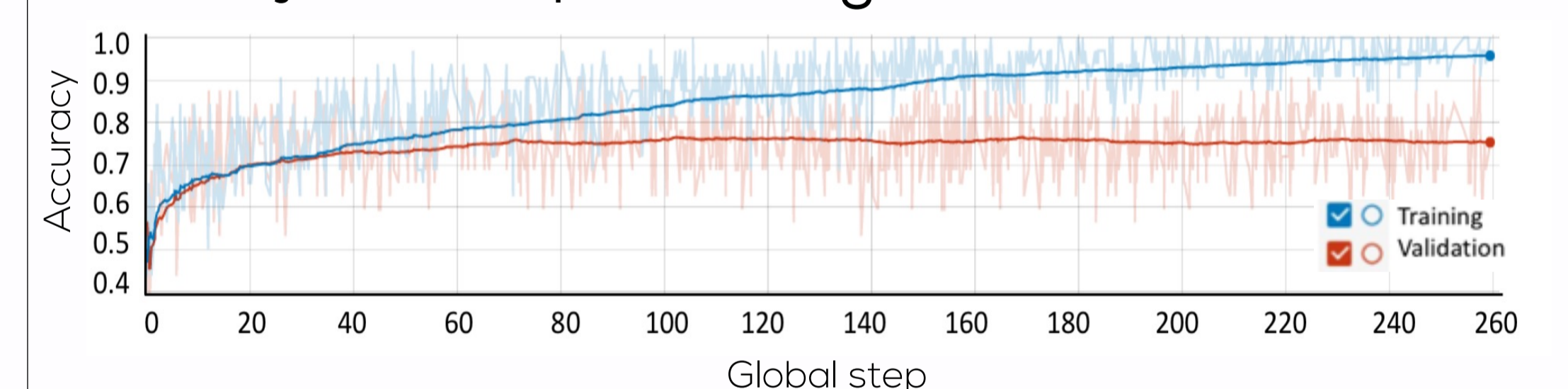


### Observation of performance bias

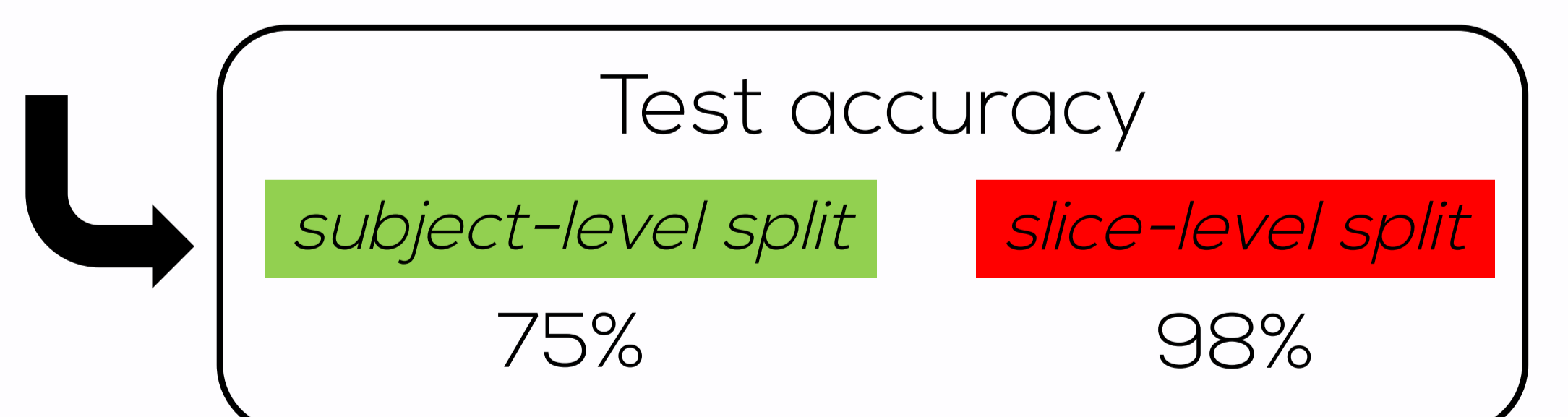
A. Slice-level partitioning



B. Subject-level partitioning



The training and validation accuracies (smoothed by a threshold of 0.99) are obtained during 150 epochs for both data split strategies over the same architecture.



## Conclusion

Data leakage is a common problem in the literature (42% of surveyed papers). Moreover, it has a serious impact on performance evaluation, as demonstrated by the strong differences in accuracies in both the literature and our experiments. Thus the current literature of the domain may **overestimate the performance of deep learning systems for automatic diagnosis** of Alzheimer's disease.

## Downloads



All the papers that were analyzed in the literature search may be found at [www.zotero.org/groups/2337160/ad-dl](http://www.zotero.org/groups/2337160/ad-dl)