Advances in Medical Sciences 64 (2019) 292–302

Contents lists available at ScienceDirect

Advances in Medical Sciences

journal homepage: www.elsevier.com/locate/advms

Original research article

Diagnostic accuracy of frontotemporal dementia. An artificial intelligence-powered study of symptoms, imaging and clinical judgement

Maksymilian A. Brzezicki[a,*], Matthew D. Kobetić[b], Sandra Neumann[c], Catherine Pennington[a]

[a] Bristol Institute of Clinical Neurosciences, University of Bristol, Southmead Hospital, Bristol, UK
[b] Faculty of Health Sciences, University of Bristol, Bristol, UK
[c] Department of Physiology and Pharmacology, Clinical Research and Imaging Centre, University of Bristol, Bristol, UK

ARTICLE INFO

Keywords:
Frontotemporal dementia
Cognitive disorders and dementia
Imaging
Computational neurology
Artificial intelligence

ABSTRACT

Purpose: Frontotemporal dementia (FTD) is a neurodegenerative disorder associated with a poor prognosis and a substantial reduction in quality of life. The rate of misdiagnosis of FTD is very high, with patients often waiting for years without a firm diagnosis. This study investigates the current state of the misdiagnosis of FTD using a novel artificial intelligence-based algorithm.

Patients & Methods: An artificial intelligence algorithm has been developed to retrospectively analyse the patient journeys of 47 individuals diagnosed with FTD (age range 52–80). The algorithm analysed the efficiency of patient pathways by utilizing a reward signal of −1 to +1 to assess the symptoms, imaging techniques, and clinical judgement in both behavioural and language variants of the disease.

Results: On average, every patient was subjected to 4.93 investigations, of which 67.4% were radiological scans. From first presentation it took on average 939 days for a firm diagnosis. The mean time between appointments was 204 days, and the average patient had their diagnosis altered 7.37 times during their journey. The algorithm proposed improvements by evaluating the interventions that resulted in a decreased reward signal to both the individual and the population as a whole.

Conclusions: The study proves that the algorithm can efficiently guide clinical practice and improve the accuracy of the diagnosis of FTD whilst making the process of auditing faster and more economically viable.

1. Introduction

Frontotemporal dementia (FTD) is a neurodegenerative disorder associated with a poor prognosis and dramatic reduction in quality of life [1]. FTD accounts for up to 20% of young-onset dementias [2] and despite recent advances in imaging techniques, codification of diagnostic criteria and epidemiological studies, the rate of misdiagnosis is still very high, particularly early on in the disease [3,4]. Because symptoms can mimic various other neurological and psychiatric diseases (as indeed shown in the 'Results' section of the present paper), FTD patients are often subjected to a number of different investigations, clinic visits, scans, and biochemical tests before the correct diagnosis can be made [5].

New treatments are emerging for the different variants of dementias, and thus an early and accurate diagnosis of dementia subtype is increasingly more important [6–8]. An accurate panel of diagnostic tests could also minimize the distress for the patient, sparing unnecessary investigations and misdiagnoses they currently experience [9]. This will also reduce associated healthcare costs by avoiding inappropriate investigations and treatments.

1.1. The variability and inaccuracy of the diagnostic criteria

FTD presents itself in two major variants: behavioural and linguistic (non-fluent and semantic), which complicates the task of standardizing the criteria. The histopathological examinations is heterogeneous in FTD cases as different symptoms have been associated with various different neuropathologies and genotypes [10].

There are no universally accepted criteria for the diagnosis of the FTD. The Neary criteria [11] recognize all three clinical phenotypes of FTD and incorporate them into a universal panel. More recently, however, Rascovsky published a more detailed criteria for diagnosing behavioural variant FTD [12] according to its likelihood (possible, probable and definite). Semantic and non-fluent variants of FTD are

* Corresponding author at: Bristol Institute of Clinical Neurosciences, University of Bristol, Southmead Hospital, Southmead Road, Westbury-on-Trym, BS10 5NB, Bristol, UK.
E-mail address: mbrzezicki@neurologicalsociety.org (M.A. Brzezicki).

https://doi.org/10.1016/j.advms.2019.03.002
Received 9 July 2018; Accepted 19 March 2019
Available online 02 April 2019

usually assessed using the Gorno-Tempini criteria [13]. Our study reports an analysis for all three groups of criteria, to maximize the clinical application of the paper's findings.

To enhance the accuracy of these panels, various improvements have been proposed. Apraxia examination [14], neuropsychometry [15,16], or amyloid-β 2–42 CSF analysis [17] have all been used to better differentiate FTD from other types of dementia. Nevertheless, when subjected to confounders such as Alzheimer's disease or Lewy Body Dementia, the criteria reveal a very poor sensitivity for FTD [18,19]. Indeed, our study reports on the most common ones in the 'Results' section below.

Some point to a better use of imaging techniques as a *panacea* for the problem. However, diagnosing FTD from an MRI scan can be very difficult, particularly in the early stages of the disease [20,21]. An addition of a multivariate analysis [22] or a PET scan may improve the accuracy [23,24], albeit only to a limited extent [25].

Single positron emission computer tomography (SPECT) emerged as a potentially useful tool, as it attempts to directly measure the perfusion of the affected areas (i.e. frontal and temporal lobes). However, there is a mixed picture in the literature concerning these scans, with their considerable limitations [26,27]. A recent meta-analysis revealed that there is currently insufficient evidence to recommend the use of SPECT scans in routine practice [28]. EEG may be helpful [29–31] but it is not usually employed in a dementia clinic.

Another difficulty is that of the disease progression [32]: the symptoms of the FTD can vary at different stages of presentation.

1.2. Machine learning and pattern recognition

The inaccuracies of human perception could potentially be mitigated by the recent developments in the field of computational biology. Several studies demonstrated that artificial intelligence (AI)-developed methods of imaging analysis and pattern recognition can reliably improve the diagnostic accuracy when compared to a human-based assessment (e.g. in classification of fractures [33], dermatological assessment [34], or MRI interpretation [35–39]).

We have expanded the scope of that analysis to include the entirety of the patient's journey: from the first presentation to neurology services to the final diagnosis of FTD.

In this paper, we present a result of a retrospective longitudinal AI-powered study of patients with FTD diagnosis, assessing the progression of their disease in time, most common misdiagnoses, and the reasons for inaccuracies. The study aims to report on the current state of the problem and to propose potential improvements in the clinical practice.

2. Patients and methods

2.1. Participant selection

The participants were selected from an anonymized database from the Memory Clinic in the Brain Centre, Southmead Hospital, Bristol, UK. This is a tertiary referral centre and a major university hospital for neurosciences in the South West of England (5.6 m population) [40]. The main secondary care service, however, is provided within the Bristol metropolitan and Somerset area. The database features the diagnosis of record, and in order to be included, a patient had to have a full diagnosis of any FTD variant. At the time of search, records included hospital visits from 2009 to 2017. The selection process is illustrated in Fig. 1.

2.2. Frideswide algorithm set-up

A total of 47 patient journeys were followed and analysed using the Frideswide AI algorithm (FwA) [41]. The input was generated from clinical letters, discharge summaries, request sheets and investigation reports available on the hospital's computer system. The date of admission (and the first point on the timeline to be analysed) was set to the first presentation with a neurological symptom, be it to the Emergency Department (ED), or via a referral from the General Practitioner (GP). The last follow-up was concluded when the patient was given a firm diagnosis by a neurologist (Fig. 2). For the secondary analysis of the FTD variants, the first point on the timeline was adjusted to the first neurology services encounter: on the ward or in clinic, since it would be unlikely for an ED physician or a GP to use criteria-based evaluation of the disease.

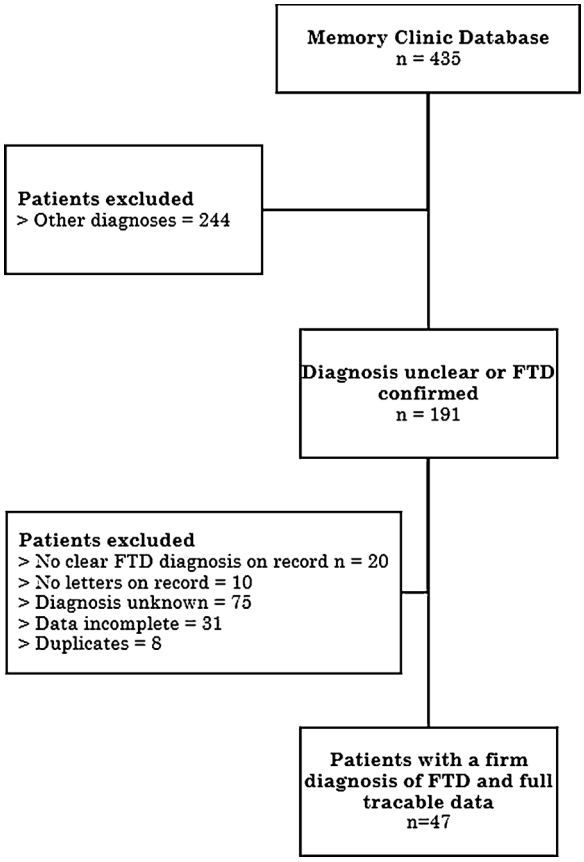

Fig. 1. Flowchart of the process of systematic review of the database and selection of the patients for analysis.

2.3. Reward assignment

Amid the healthcare informatics revolution, the "AI" and "machine learning" terms have been used in a variety of different situations [42]. For clarity, the FwA is an analytic tool that acts as an *intelligent agent* [43]. It studies the environment composed of the available data (clinical information from letters, reports, requests, etc.) in different time frame nodes, each healthcare encounter being a different node. It then attempts to achieve the best maximisation of the diagnostic efficiency (*goal function* in AI nomenclature), i.e. so that the patient is diagnosed at a minimum time, cost and invasiveness.

The AI uses the doctor-defined pay-off table to determine the values of different decisions it may take. The overall aim is to produce a list of improvements that could achieve the most optimal efficiency for the entire hospital. The exact process is described below and in Fig. 2.

After the data was imported, each visit to the hospital was represented as a point on the timeline, with corresponding symptoms, current differential diagnoses, care plans and actions. The appointments were then associated with each other on an action-outcome basis, in a chronological order. This data model was then enhanced by adding information about diagnostic accuracy, time, cost, and invasiveness of

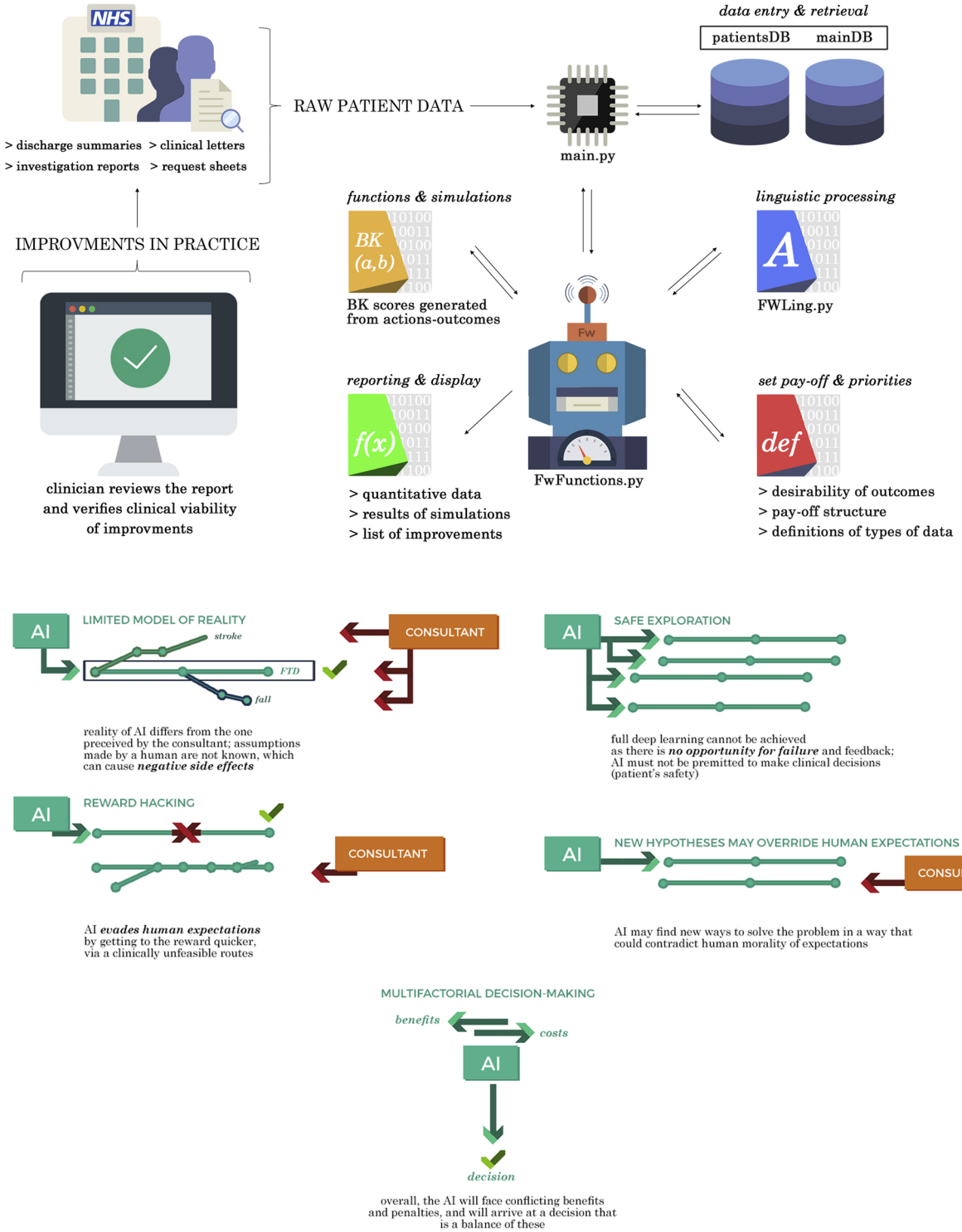

Fig. 2. (A) The computational construct and system of the simulation. Physical coding files are marked with an extension ".py" (Python), the databases are marked with a suffix "DB" (Sqlite). (B) Structure of the basal (obligatory) data model and decision network of analysis used by the algorithm Different behaviors exhibited by the algorithm during the study.

the interventions (Fig. 2).

The *utility function* [44] of the FwA assigned the BK coefficients to each of the action-outcome associations. A BK coefficient is a score that helps the AI understand whether an outcome was clinically desirable. It translates what a human being would consider a good or a bad outcome into a numerical figure that the AI can understand. This way, the AI was able to evaluate an association, by having a set of standards and values. Each association was assigned a *pay-off award* in the form of a BK coefficient, which represents an association that is desirable clinically on the scale of 1 to -1, from the most to the least desired outcome.

This was dictated by the following payoff set: {firm diagnosis: 1,? diagnosis: 0.5, very close diagnosis: 0.3, associated diagnosis: 0.25, unknown or entirely different diagnosis: –1}. A "?" sign was used to signalize that a diagnosis was not certain at a given point in time.

This way, the action that resulted in the firm diagnosis of FTD would be ideal (BK = 1), and the action that resulted in a firm diagnosis of a

Table 1
Summary of the quantitative panels executed in the analysis of Frontotemporal Dementia patients included in the study.

Demographics	Behavioural variant	Language variant (non-fluent, semantic)	
N	26	21	
Mean age on diagnosis (range), ± SD	67.25 (52-83) ± 8.99	68.09 (54-80) ± 9.10	
M:F ratio	3:2	1:1	
Neuropsychometry			
Mean MoCA on diagnosis (range), ± SD	6.96/30 (2-24) ± 8.20	5.48/30 (2-24) ± 7.17	
Provision of healthcare *(from the first visit in neurology services)*			
Mean number of times the diagnosis of record was changed per patient (range) ± SD	8.31 (3-25) ± 4.82	6.19 (3-10) ± 2.32	$t_{36} = 1.93$ p = 0.06
Mean time between presentation to diagnosis (range) ± SD	916 days (1-2607) ± 774	795 days (84-2478) ± 629	$t_{45} = 0.59$ p = 0.56;
Mean time between various hospital appointments (range) ± SD	99 days (17-434) ± 181	233 days (41-877) ± 198	$t_{28} = 1.11$; p = 0.277
Healthcare interventions (mean per patient)	12.15	10.43	
Investigations total	5.31	3.95	
Radiological investigations	3.42	2.71	
Referrals (excluding initial referrals)	0.31	0.52	
Follow up and wait	0.81	0.52	

different disease, e.g. Huntington's disease would be least desirable (BK = – 1)

The algorithm then attempted to approximate the state of the world [45], i.e. the neurology service provided for the patients as analysed in the study, to the ideal overall BK coefficient of BK = 1.

To that end, it suggested a number of *improvements*, which were collected in a matrix. An *improvement* was a FwA-generated suggestion of how the service can be enhanced. In this context, an *improvement* is something that would enhance the diagnostic accuracy. These improvements were then tested against their feasibility across the entire population, to see whether they would work if they were actually introduced as a routine service element to be applied to all patients.

2.4. Improvement analysis and reporting

If, by removing the negative element from the journey, the entire population were better off (overall BK increased), the improvement was deemed significant. If, on the contrary, the negative element's contribution was beneficial to the large sample overall, it was deemed a *necessary evil*, as enough patients in the population benefited from that element to warrant its usefulness.

2.5. If it can diagnose, it can misdiagnose

Some investigations are performed to exclude an alternative diagnosis or to not to miss a sinister or less common aetiology. Therefore, to correct for this phenomenon, only investigations which at any point reported a diagnosis of FTD were subsequently included in the analysis. Those which did not purport to be capable of making that diagnosis were not evaluated for that purpose.

2.6. Natural history of the disease

A battery of quantitative panels was also performed to assess the state of the FTD population. Since the study collected a large number of symptoms, imaging reports and investigations results, we have reported on the natural history of the FTD development, as captured during various hospital appointments and examinations.

Notwithstanding the variability of the available diagnostic tools, we have decided to evaluate the accuracy of the FTD diagnosis with the most widely used criteria [11]. These include measurements of behavioural change, linguistic problems, pattern of social and personality changes over time and supportive information from neuropsychometry, EEG, and imagining. One point was awarded for every element on the panel (both sections I and II of the criteria [11]).

The patients were then sub-divided into language (both non-fluent primary progressive aphasia and semantic form) and behaviour variants of the disease, and further assessed using the Raskovsky and Gorno-Tempini criteria, respectively [12,13]. The assessment was used to evaluate any differences in care between the two variants and to report a natural history of the disease over time.

Some participants were offered neuropsychometry as well as genetic tests. However, because of the computer system set-up, the dates of these appointments cannot be reliably established. Nevertheless, they have still been included in the analysis. Their date was set at the time of the patient learning about the outcome, which was the date of the next appointment.

2.7. Ethical issues

The project was reviewed and received a favourable ethical opinion from NHS Health Research Authority, IRAS project ID: 209781.

2.8. Statistical analysis

Descriptive statistics were used to summarize the quantitative sections. Overall effect of the improvements was calculated in a meta-analysis statistical panel, including z-test for total fixed effects, and Cochrane's Q and I^2 tests for heterogeneity. We appreciate that these statistical tools are commonly associated with a meta-analysis of studies; in this context, the independence assumption is not valid. Thus, the metrics are intended to aid the understanding of the decision-making process and to confirm whether an AI recommendation was statistically significant, i.e. whether a phenomenon indicated in the results could be explained by the data used for its evaluation. The odds ratio (OR) and ANOVA were used in secondary analysis of the power of a request form. Unpaired two-tailed *t*-test was used to evaluate the differences between variants of FTD. The statistical significance level was set for the $p < 0.05$.

3. Results

3.1. Quantitative panel

The demographics, times and a categorized breakdown of symptoms and healthcare services received by the patients are presented in Table 1. Before reaching the diagnosis, on average, every patient was subjected to 4.93 distinct types of investigations, 67.4% of which were radiological scans (CT, MRI, ^{123}I-ioflupane nuclear neuroimaging - DaTscan, and SPECT).

From presenting to the hospital with a neurological symptom, be it from a GP referral or via the emergency department, an average patient waited 939 days (just under 2 years and 7 months, range 18–2911 ± 758 days) for their firm diagnosis. The mean time between

the various hospital appointments (including scans) was 204 days (6 months 3 weeks, range 17–877 ± 150 days). The average patient had their diagnosis of record changed 7.37 times (range 3–25 ± 4) during their journey. There was no statistical difference between behavioural and language variants of FTD in any of these measurements. The full breakdown is displayed in Table 1.

3.2. Diagnostic criteria

Globally, the mean diagnostic Neary score at the first presentation was (out of 24): 0.89 ± 1.81 (range 0–8), and the mean culminated score at the time of a firm diagnosis was 9.78 ± 5.72 (range 2–28). The progression of the mean culminated score over the course of hospital appointments is illustrated in Fig. 5A, with sensitivities for a clinical diagnosis with different cut-off scores is shown in Fig. 5B.

The most common categories of symptoms are presented with their prevalence on the first and subsequent appointments, along with their overall prevalence throughout the journey in Fig. 5C. The progression of the disease in behavioural and language variants of the disease is illustrated in Fig. 6.

3.3. Conducting differential diagnosis

Whilst some patients received their initial indication of FTD fairly promptly (in less than 9 months) after presenting with their symptoms, others were misdiagnosed or undiagnosed for a long time (up to 2911 days). Out of the latter group, one pattern of differential diagnosis featured a series of confounding diagnoses initially (Fig. 3), and then a narrowing-down process leading to the FTD diagnosis.

The other pattern consisted of a series of negative pivotal points, whereby an indicative diagnosis of FTD was removed from consideration and replaced by either a confounding diagnosis or lack of any diagnosis at all. The most common change generating a negative BK score was a suspicion of the FTD to unknown (18 cases), other types of dementia to unknown (11 cases) and a suspicion of FTD into other types of dementia (6 cases).

The breakdown of negative and positive BK results of particular interventions is presented in Fig. 4 Fig6. The total time for computing and analysing of the aforementioned data was 13.95 s (Fig. 7).

4. Discussion

Overall, this work shows that the best and most accurate diagnosis of FTD is still made clinically. The usefulness of radiological investigations ranged from borderline helpful to being a substantial distractor, especially when used not to *exclude* an alternative diagnosis but to *confirm* the FTD (Fig. 4).

Based on the obtained data, the large waiting times and stress of coming to the hospital could have been substantially alleviated if the diagnosis had been made based on clinical picture and biochemical blood investigations only, to exclude other diagnoses (i.e. organic causes of the symptoms, such as electrolyte imbalance, liver dysfunction, autoimmune and paraneoplastic disorders, etc.).

The benefits of performing the MRI or SPECT scans were demonstrated to be outweighed by the costs, time, and distress to the patient, especially in the case of the MR imaging.

It also must be noted that the invasiveness of the scan may subjectively be more significant to a patient experiencing FTD-like symptoms, than to a healthy control.

However, the value of the actual time saved is difficult to establish. Frequently, the follow-up appointment will only happen once the test planned from the previous appointment is done. Furthermore, the progression of the disease in time could be important diagnostically, as several neurodegenerative disorders present with distinct natural histories.

CT scans reported numerous BK = 0 neutral scores, as their use was, in many cases, to exclude other diagnoses or more acute syndromes.

4.1. Reading letters saves lives

Interestingly, inability to read the clinical letters was one of the major cause of negative BK scores in the diagnostic process of the FTD. Lack of proper familiarization with the current patient's notes resulted in a substantial number of unnecessary referrals and investigations that either did not contribute any relevant clinical information or, even worse, distracted from the FTD diagnostic pathway.

One reason for this phenomenon could be a lack of an integrated clinical letter system in Southmead Hospital, as indeed is seen in many other hospitals. For instance, Emergency Department clinicians may not be aware of clinical letters from the Neurology Department and could thus *bona fide* assume that the current admission is the patient's first presentation with the symptoms. This may be aggravated by the difficulty of obtaining a detailed history from a patient with memory and language problems.

The caregivers of the patient attending the hospital may not have the copies of clinical letters at hand. For an assessing physician, the collateral history may offer a limited degree of reassurance. Thus, they may be requesting additional scans to confirm the diagnosis of a neurodegenerative disorder. We appreciate there may be many further reasons for not being able to read the cognitive neurology letters.

4.2. The art and power of a request form

Whilst the SPECT scan is widely requested for the diagnosis of FTD, the clinical evidence for its use is still lacking [28]. This type of investigation was the only FTD-specific non-clinical benchmark identified

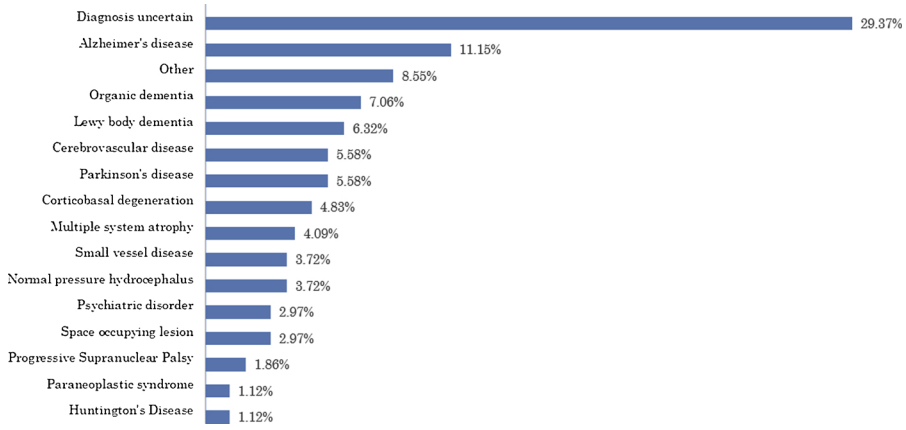

Fig. 3. Quantitative summary of the different diagnoses assigned to patients on the course of their journeys.

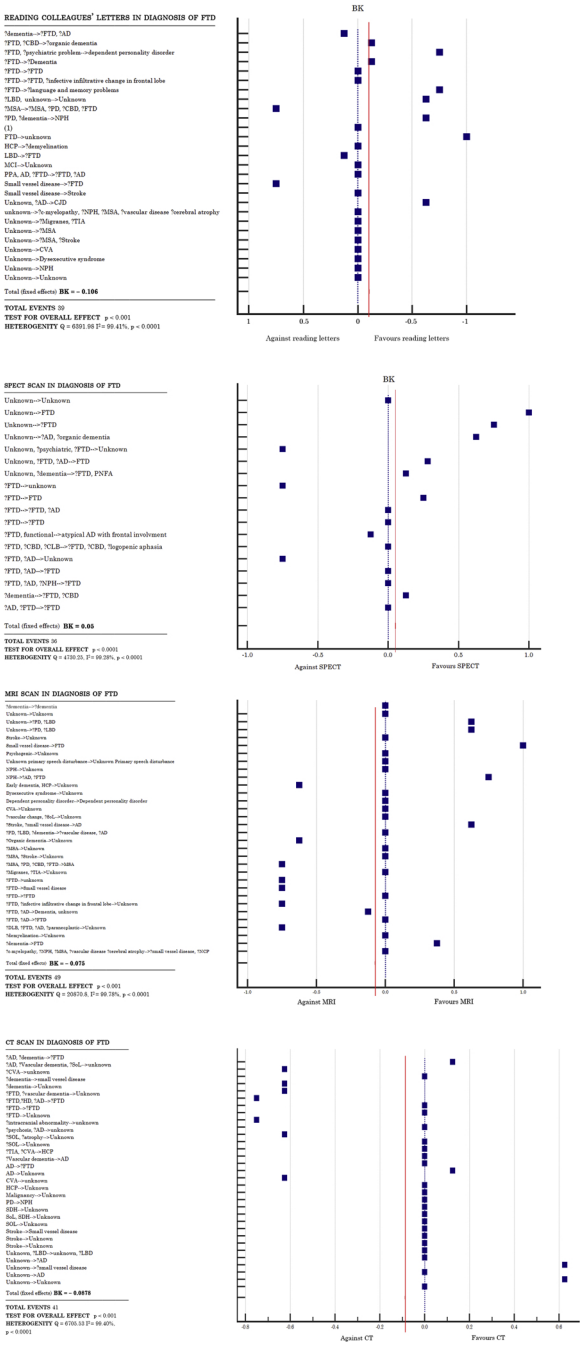

Fig. 4. Summary of BK coefficients for different interventions in diagnostic accuracy of frontotemporal dementia (FTD). A red line indicates the point of overall effect. AD = Alzheimer's Disease, PNFA = Progressive non-fluent aphasia, CBD = Corticobasal degeneration, NCP, NPH = Normal pressure hydrocephalus, PD = Parkinson's Disease, LBD = Lewy Body Dementia, CVA = Cerebrovascular event, SoL = Space occupying lesion, MSA = Multisystem atrophy, TIA = Transient ischemic attack, HD = Huntington's disease, SDH = Subdural hemorrhage, MCI = Mild Cognitive Impairment. (1) = ? psychiatric disorder,? neurodegenerative dementia,? FTD,? paraneoplastic –> ?DLB,? FTD,? AD,? paraneoplastic.

in the study and its overall performance was positive, albeit with a substantial number of occasions where its negative findings severely delayed the diagnosis.

Secondary analysis of the reports and requests for the SPECT scan revealed a very interesting relation: by putting the suspicion of the FTD on the request form, there was an OR = 5.75 (z = 2.101, p = 0.036) of receiving a positive report (i.e. one reporting that the scan was in keeping with FTD), whereas a negative finding was independent of a request form. Not putting the FTD on the form resulted in a random chance of getting an either positive or a negative report. There was no difference ($F_3 = 0.446$; p = 0.72) between the groups in terms of symptoms or clinically-assessed progression of the disease.

This may either indicate a confirmation bias in clinicians over-diagnosing the FTD after a positive scan or a more diligent scan interpretation prompted by the suggestion on the form.

4.3. Towards an accurate diagnosis

There have been several attempts to assess the sensitivity of the FTD scoring panels in the past: 79% [46] and 36.5% [47] were reported for the Neary criteria. These figures would be replicated in our study by applying a cut-off point of 5.

Interestingly, we have been able to demonstrate how the mean culminated score on the Neary criteria is changing over time, which allows for the application of a cut-off point not only in terms of desired sensitivity (relative to clinical diagnosis), but also within a specified timeframe.

We have, however, deliberately exercised caution in reporting the sensitivity figures, as they may not be very reliable in the absence of a gold standard independent of the consensus criteria. Specificity was also not available because of ethics constraints on the current study, disallowing the inclusion of non-FTD cases.

Per analogiam, we have refrained from proposing our own adjustments to the diagnostic criteria at this stage.

4.4. Concrete problems of AI implementation

Putting the FwA to a real-life test revealed several phenomena that were not displayed during the initial simulation [41] (Fig. 2A). Interestingly, these were hypothesized in the landmark AI paper before [48] and thus the study is of an immense value to the computer scientists exploring this unchartered territory.

4.4.1. Avoiding negative side effects

Since the model of reality presented to the algorithm differs from the one perceived by the consultant, it is difficult to prescribe all the *negative side effects* [48] that a neurologist would have taken for granted, e.g. an acute confusion and slurred speech warrant a CT scan with suspected cerebrovascular event. The algorithm, however, perceives it as a deviation from its goal and is willing to sacrifice the patient's health, ignore the stroke, and allow the ischemia to go on, only to get to the FTD diagnosis quicker. We have alleviated the impact of this phenomenon with a supervised learning of these side effects, yet we appreciate that there is a vast number of potential *negative* and *positive* side effects that are still unaccounted for.

4.4.2. Avoiding reward hacking

The FwA quickly acquired a skill of evading our intended outcomes. It managed to accelerate the process of getting to the reward quicker by generating improvements that would not be clinically feasible.

One of these phenomena was a *genius effect*, whereby by removing a confounding factor, e.g. a test result, the consultant in the latter part of the association appeared to have made the diagnosis out of their skill, whereas in fact, they would have been aware of the result that was artificially removed.

We managed to largely eradicate this by introducing more distinct, sometimes conflicting dimensions contributing to the BK coefficient. This made the algorithm consider several factors when making a decision, a situation similar to the one faced by a clinician.

We anticipate that further sophistication of the coefficient would not only make the result more clinically applicable but will contribute to minimizing the reward hacking.

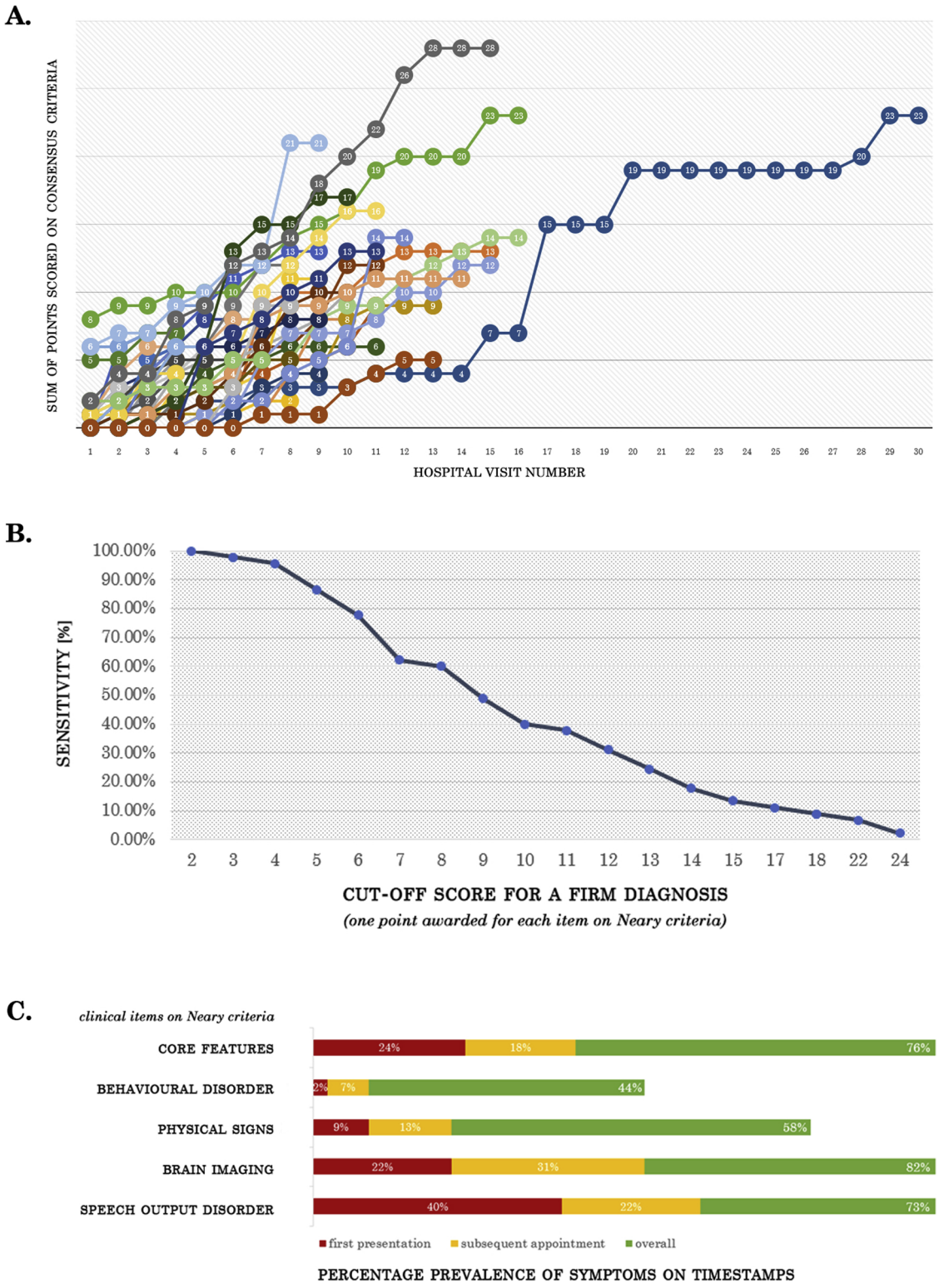

Fig. 5. (A) Trace of sum of points scored on the consensus diagnostic criteria for Frontotemporal Dementia (FTD) over time across various hospital appointments. Once point is awarded for every item on the panel (both sections I and II). The last entry for each trace is the time of firm FTD diagnosis (B) Sensitivity scores for different cut-off points applied at the time of the hospital appointment with a firm FTD diagnosis (C) Prevalence of the core diagnostic features on different stage of patient's journey, as per *Neary* criteria: Core diagnostic features (incl. insidious onset, impairment of personal and interpersonal conduct regulation); Behavioral disorder (incl. decline in hygiene, mental rigidity, hyperorality, perseverative behavior); Physical signs (incl. akinesia, rigidity, tremor, balance problems); Speech and language deficits (incl. altered speech output, mutism).

4.4.3. The scalable oversight and safe exploration

These problems were avoided by a sheer volume of data which did not require separate supervised teaching exercises. Furthermore, the algorithm was not allowed to make decisions about the treatment or diagnosis at this stage.

However, an argument could be put forward that the sandbox

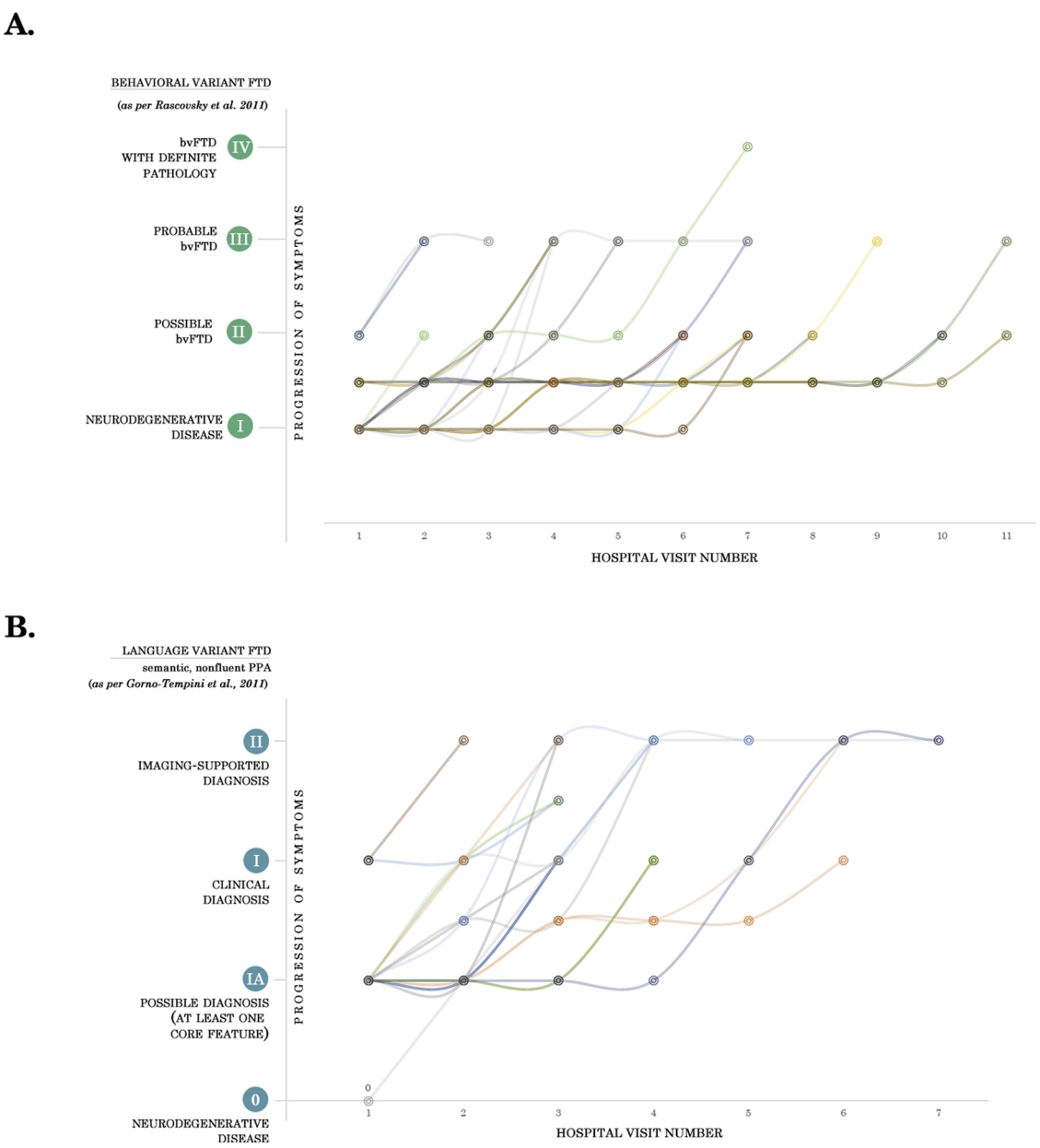

Fig. 6. (A) Trace of the progression of symptoms in behavioral variant FTD (B) and language variant FTD: both semantic and non-fluent primary progressive aphasia (PPA). Roman numerical signs correspond to the levels of confidence described in respective diagnostic criteria.

simulation omitted the unexpected consequences of these decisions. This is inevitable in clinical practice, where a great deal of caution must be exercised when trailing the AI systems. We envisage that a slow and methodological implementation of the suggestions and a re-audit cycle would be a safe approach to this problem, as is in the case of a standard clinical audit.

4.5. The advantage of AI use

The data analysis and results presented in this report could have been achieved by standard statistics and by using auditing methods currently available. However, the AI use offers certain advantages.

First, the superiority of the FwA is in the time saved in performing the analysis. The overall process took 24.5 s on a standard PC machine (Intel ® Core ™ i7-4790 CPU @ 3.60 GHz, 16.0 GB RAM, Windows 10 64-bit, x64-based processor). During that, the FwA analysed 1,597 clinical findings and 515 clinical decisions, generated 44 hypotheses and conducted 249 experiments, where it re-evaluated the patients' journeys, back to back, to test whether a hypothesis is clinically feasible. This volume of data would require considerably more time and resources to be analysed by human means.

Furthermore, a classic audit would usually involve testing a hypothesis *that is known* to the researcher before starting the analysis. The FwA not only has the capacity to test multiple hypotheses without a significant increase of time required but can also generate new hypotheses on its own. This means that the FwA-generated audit can reveal a clinically valuable answer to the question that is not known *a priori*.

It is also possible to adapt the FwA for other clinical scenarios e.g. time taken to diagnose lung cancer. The same program can perform multiple different analyses across different specialties.

4.6. Limitations of the study

The design of the study was aimed at analysing the FTD population only. Hence, the FwA had a limited impact of confounding diagnoses or

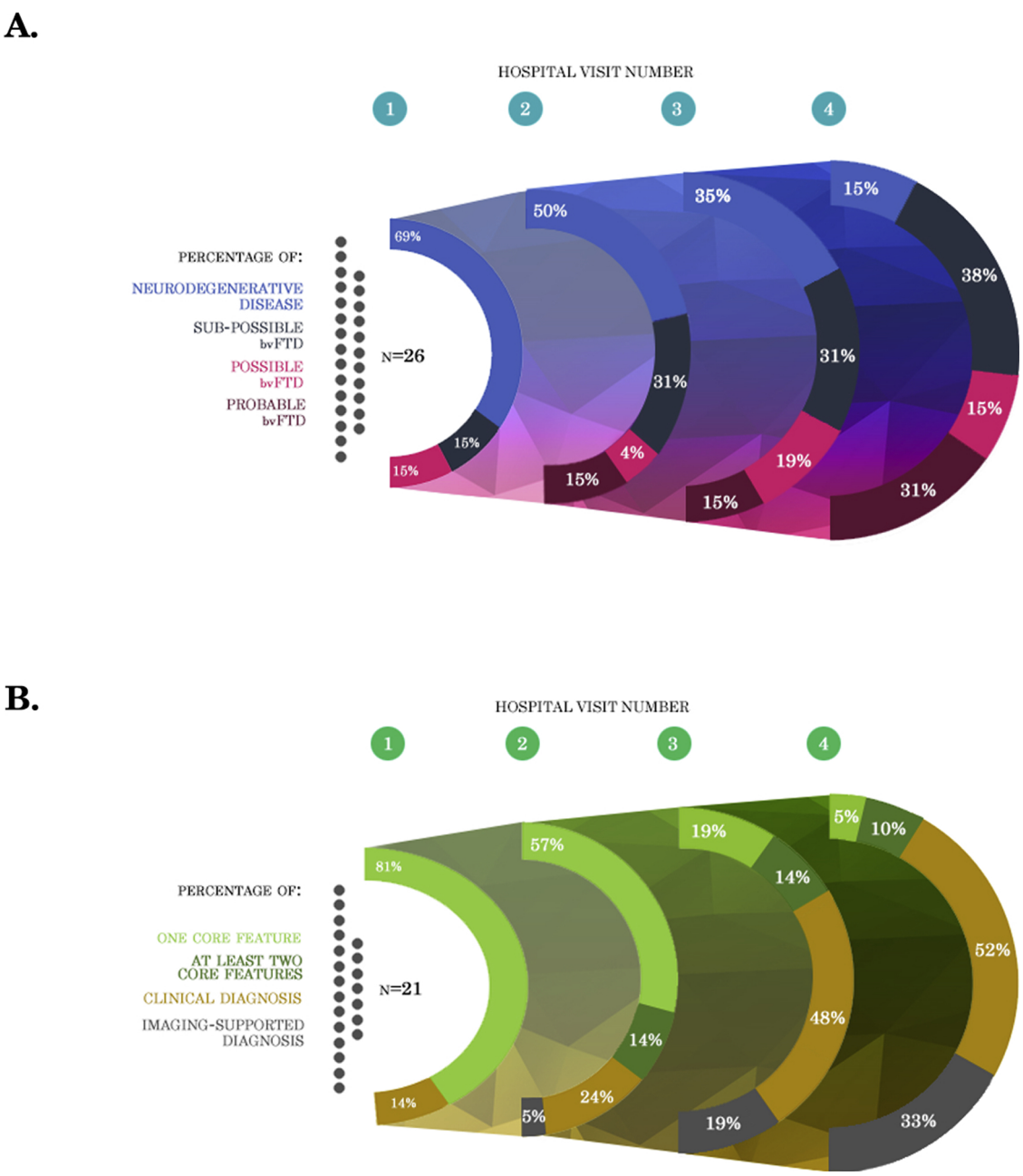

Fig. 7. (A) Progression of symptoms in behavioral variant FTD and (B) language variant FTD The percentage points are numbers of patients diagnosed with given clinical certainty during a respective hospital appointment.

indeed the journeys whereby the FTD was mentioned as a differential and an alternative firm diagnosis was made at the end. Thus, the improvements suggested by the algorithm can enhance the sensitivity but not specificity of the FTD diagnosis.

Furthermore, since the firm diagnosis of FTD was made by the consultant, there is a limited space to scrutinize the clinical assessment efficacy. To alleviate this, an alternative gold standard could be introduced, e.g. a histopathological examination, which would provide an independent indicator of accuracy. We appreciate that this can limit the applicability of the statistical results; this is a common problem in diseases that cannot be definitely diagnosed *ante mortem*.

We appreciate that there may be further difference to be elucidated in non-fluent vs. semantic variants of the language FTD. However, the study was not powered to investigate these in detail and thus the groups were displayed together in the figures and tables. Nevertheless, the Gorno-Tempini diagnostic criteria have been applied appropriately to the respective variants.

The study presents the analysis of the population under care of a major neurology centre in the South West of England. We appreciate that a multi-centre study is warranted for a better external validity.

Furthermore, since the FTD incidence is 2.4–4.1 per 100,000 [1], it is difficult to obtain a large dataset of patients for the analysis. This is a major limitation in the generalisability of the FwA. However, large data requirement can limit many smaller audits in clinical practice. This paper can serve as an indication that smaller studies, e.g. on rare diseases or on a small population are still feasible.

5. Conclusions

Frontotemporal dementia is still a very difficult to diagnose condition. Patients often wait for years to receive their diagnosis, and during that time they are subjected to a number of potentially distressing and

sometimes unnecessary investigations. They are often given different disease names to explain their memory and language problems, and each of those is associated with a shock and stress of learning about the condition. This comes at a cost to the patient's psychological wellbeing and puts an extra pressure on the already stretched healthcare service resources.

Since FTD is associated with a shorter life expectancy and poorer prognosis than more conventional dementias, an accurate and prompt diagnosis is of a paramount importance in planning of treatment and making personal life choices by the patient.

This can be achieved by a constant audit cycle, implementation of improvements in diagnostic technique and a repeated evaluation of patient journeys. Normally, that process would require extensive amount of effort, funds, and staff hours.

This study, however, proves that by using an AI-powered algorithm, the process could be made more economically viable and could easily produce the results to guide a better clinical practice and to significantly improve the patients' experience in a cognitive neurology clinic.

Acknowledgements

The authors would like to thank Dr Liz Coulthard, Consultant Senior Lecturer in Dementia Neurology at the University of Bristol for her support and inspiration for the future development of the algorithm within the clinical neurosciences.

We also would like to thank Robert Miles, an AI scholar at the University of Nottingham for his invaluable expertise that allowed this paper to take its current shape.

Conflict of interest

The authors declare no conflicts of interests.

Financial disclosure

The authors have no funding to disclose.

Author contributions

Study Design: Maksymilian A. Brzezicki, Catherine Pennington, Matthew D. Kobetic, Sandra Neumann

Data Collection: Maksymilian A. Brzezicki, Catherine Pennington

Statistical Analysis: Maksymilian A. Brzezicki, Catherine Pennington, Matthew D. Kobetic, Sandra Neumann

Data Interpretation: Maksymilian A. Brzezicki, Catherine Pennington, Matthew D. Kobetic, Sandra Neumann

Manuscript Preparation: Maksymilian A. Brzezicki, Catherine Pennington, Matthew D. Kobetic, Sandra Neumann

Literature Search: Maksymilian A. Brzezicki

Funds Collection: N/A

References

[1] Onyike CU, Diehl-Schmid J. The epidemiology of frontotemporal dementia. Int Rev Psychiatry Abingdon Engl 2013;25:130–7. https://doi.org/10.3109/09540261.2013.776523.

[2] Snowden JS, Neary D, Mann DMA. Frontotemporal dementia. Br J Psychiatry J Ment Sci 2002;180:140–3.

[3] Qiu C, Kivipelto M, von Strauss E. Epidemiology of Alzheimer's disease: occurrence, determinants, and strategies toward intervention. Dialogues Clin Neurosci 2009;11:111–28.

[4] Shinagawa S, Catindig JA, Block NR, Miller BL, Rankin KP. When a little knowledge can Be dangerous: false-positive diagnosis of behavioral variant frontotemporal dementia among community clinicians. Dement Geriatr Cogn Disord 2016;41:99–108. https://doi.org/10.1159/000438454.

[5] Landqvist Waldö M, Gustafson L, Passant U, Englund E. Psychotic symptoms in frontotemporal dementia: a diagnostic dilemma? Int Psychogeriatr 2015;27:531–9. https://doi.org/10.1017/S1041610214002580.

[6] Seeley WW. Frontotemporal dementia neuroimaging: a guide for clinicians. Front Neurol Neurosci 2009;24:160–7. https://doi.org/10.1159/000197895.

[7] Pasquier F. Telling the difference between frontotemporal dementia and Alzheimer's disease. Curr Opin Psychiatry 2005;18:628–32. https://doi.org/10.1097/01.yco.0000185988.05741.2a.

[8] Duara R, Barker W, Luis CA. Frontotemporal dementia and Alzheimer's disease: differential diagnosis. Dement Geriatr Cogn Disord 1999;10(Suppl 1):37–42. https://doi.org/10.1159/000051210.

[9] Bang J, Spina S, Miller BL. Frontotemporal dementia. Lancet Lond Engl 2015;386:1672–82. https://doi.org/10.1016/S0140-6736(15)00461-4.

[10] Sieben A, Van Langenhove T, Engelborghs S, Martin J-J, Boon P, Cras P, et al. The genetics and neuropathology of frontotemporal lobar degeneration. Acta Neuropathol (Berl) 2012;124:353–72. https://doi.org/10.1007/s00401-012-1029-x.

[11] Neary D, Snowden JS, Gustafson L, Passant U, Stuss D, Black S, et al. Frontotemporal lobar degeneration: a consensus on clinical diagnostic criteria. Neurology 1998;51:1546–54.

[12] Rascovsky K, Hodges JR, Knopman D, Mendez MF, Kramer JH, Neuhaus J, et al. Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia. Brain 2011;134:2456–77. https://doi.org/10.1093/brain/awr179.

[13] Gorno-Tempini ML, Hillis AE, Weintraub S, Kertesz A, Mendez M, Cappa SF, et al. Classification of primary progressive aphasia and its variants. Neurology 2011;76:1006–14. https://doi.org/10.1212/WNL.0b013e31821103e6.

[14] Johnen A, Tokaj A, Kirschner A, Wiendl H, Lueg G, Duning T, et al. Apraxia profile differentiates behavioural variant frontotemporal from Alzheimer's dementia in mild disease stages. J Neurol Neurosurg Psychiatry 2015;86:809–15. https://doi.org/10.1136/jnnp-2014-308773.

[15] Lemos R, Duro D, Simões MR, Santana I. The free and cued selective reminding test distinguishes frontotemporal dementia from Alzheimer's disease. Arch Clin Neuropsychol Off J Natl Acad Neuropsychol 2014;29:670–9. https://doi.org/10.1093/arclin/acu031.

[16] Beck IR, Schmid NS, Berres M, Monsch AU. Establishing robust cognitive dimensions for characterization and differentiation of patients with Alzheimer's disease, mild cognitive impairment, frontotemporal dementia and depression. Int J Geriatr Psychiatry 2014;29:624–34. https://doi.org/10.1002/gps.4045.

[17] Bibl M, Gallus M, Welge V, Esselmann H, Wolf S, Rüther E, et al. Cerebrospinal fluid amyloid-β 2-42 is decreased in Alzheimer's, but not in frontotemporal dementia. J Neural Transm Vienna Austria 1996;2012(119):805–13. https://doi.org/10.1007/s00702-012-0801-3.

[18] Harris JM, Thompson JC, Gall C, Richardson AMT, Neary D, du Plessis D, et al. Do NIA-AA criteria distinguish Alzheimer's disease from frontotemporal dementia? Alzheimers Dement J Alzheimers Assoc 2015;11:207–15. https://doi.org/10.1016/j.jalz.2014.04.516.

[19] Schmid NS, Ehrensperger MM, Berres M, Beck IR, Monsch AU. The extension of the german CERAD neuropsychological assessment battery with tests assessing subcortical, executive and frontal functions improves accuracy in dementia diagnosis. Dement Geriatr Cogn Disord Extra 2014;4:322–34. https://doi.org/10.1159/000357774.

[20] Chow TW, Gao F, Links KA, Ween JE, Tang-Wai DF, Ramirez J, et al. Visual rating versus volumetry to detect frontotemporal dementia. Dement Geriatr Cogn Disord 2011;31:371–8. https://doi.org/10.1159/000328415.

[21] Koikkalainen J, Rhodius-Meester H, Tolonen A, Barkhof F, Tijms B, Lemstra AW, et al. Differential diagnosis of neurodegenerative diseases using structural MRI data. Neuroimage Clin 2016;11:435–49. https://doi.org/10.1016/j.nicl.2016.02.019.

[22] Canu E, Agosta F, Mandic-Stojmenovic G, Stojković T, Stefanova E, Inuggi A, et al. Multiparametric MRI to distinguish early onset Alzheimer's disease and behavioural variant of frontotemporal dementia. Neuroimage Clin 2017;15:428–38. https://doi.org/10.1016/j.nicl.2017.05.018.

[23] Dukart J, Mueller K, Horstmann A, Barthel H, Möller HE, Villringer A, et al. Combined evaluation of FDG-PET and MRI improves detection and differentiation of dementia. PLoS One 2011;6:e18111. https://doi.org/10.1371/journal.pone.0018111.

[24] Vijverberg EGB, Wattjes MP, Dols A, Krudop WA, Möller C, Peters A, et al. Diagnostic accuracy of MRI and additional [18F]FDG-PET for behavioral variant frontotemporal dementia in patients with late onset behavioral changes. J Alzheimers Dis JAD 2016;53:1287–97. https://doi.org/10.3233/JAD-160285.

[25] Womack KB, Diaz-Arrastia R, Aizenstein HJ, Arnold SE, Barbas NR, Boeve BF, et al. Temporoparietal hypometabolism in frontotemporal lobar degeneration and associated imaging diagnostic errors. Arch Neurol 2011;68:329–37. https://doi.org/10.1001/archneurol.2010.295.

[26] McNeill R, Sare GM, Manoharan M, Testa HJ, Mann DMA, Neary D, et al. Accuracy of single-photon emission computed tomography in differentiating frontotemporal dementia from Alzheimer's disease. J Neurol Neurosurg Psychiatry 2007;78:350–5. https://doi.org/10.1136/jnnp.2006.106054.

[27] Spehl TS, Frings L, Hellwig S, Weiller C, Hüll M, Meyer PT, et al. Role of semi-quantitative assessment of regional binding potential in 123I-FP-CIT SPECT for the differentiation of frontotemporal dementia, dementia with Lewy bodies, and Alzheimer's dementia. Clin Nucl Med 2015;40:e27–33. https://doi.org/10.1097/RLU.0000000000000554.

[28] Archer HA, Smailagic N, John C, Holmes RB, Takwoingi Y, Coulthard EJ, et al. Regional cerebral blood flow single photon emission computed tomography for detection of Frontotemporal dementia in people with suspected dementia. Cochrane Database Syst Rev 2015:CD010896. https://doi.org/10.1002/14651858.CD010896.pub2.

[29] Yener GG, Leuchter AF, Jenden D, Read SL, Cummings JL, Miller BL. Quantitative

EEG in frontotemporal dementia. Clin EEG Electroencephalogr 1996;27:61–8.
[30] Garn H, Coronel C, Waser M, Caravias G, Ransmayr G. Differential diagnosis between patients with probable Alzheimer's disease, Parkinson's disease dementia, or dementia with Lewy bodies and frontotemporal dementia, behavioral variant, using quantitative electroencephalographic features. J Neural Transm Vienna Austria 1996;2017(124):569–81. https://doi.org/10.1007/s00702-017-1699-6.
[31] Goossens J, Laton J, Van Schependom J, Gielen J, Struyfs H, Van Mossevelde S, et al. EEG dominant frequency peak differentiates between alzheimer's disease and frontotemporal lobar degeneration. J Alzheimers Dis JAD 2017;55:53–8. https://doi.org/10.3233/JAD-160188.
[32] Mendez MF. The accuracy of clinical criteria for the diagnosis of frontotemporal dementia. Int J Psychiatry Med 2004;34:125–30. https://doi.org/10.2190/8V1X-11UN-YDXR-A9KF.
[33] Chung SW, Han SS, Lee JW, Oh K-S, Kim NR, Yoon JP, et al. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta Orthop 2018:1–6. https://doi.org/10.1080/17453674.2018.1453714.
[34] Han SS, Park GH, Lim W, Kim MS, Na JI, Park I, et al. Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis: automatic construction of onychomycosis datasets by region-based convolutional deep neural network. PLoS One 2018;13:e0191493. https://doi.org/10.1371/journal.pone.0191493.
[35] Harper L, Fumagalli GG, Barkhof F, Scheltens P, O'Brien JT, Bouwman F, et al. MRI visual rating scales in the diagnosis of dementia: evaluation in 184 post-mortem confirmed cases. Brain J Neurol 2016;139:1211–25. https://doi.org/10.1093/brain/aww005.
[36] Klöppel S, Stonnington CM, Barnes J, Chen F, Chu C, Good CD, et al. Accuracy of dementia diagnosis: a direct comparison between radiologists and a computerized method. Brain J Neurol 2008;131:2969–74. https://doi.org/10.1093/brain/awn239.
[37] Ishii K, Kanda T, Uemura T, Miyamoto N, Yoshikawa T, Shimada K, et al. Computer-assisted diagnostic system for neurodegenerative dementia using brain SPECT and 3D-SSP. Eur J Nucl Med Mol Imaging 2009;36:831–40. https://doi.org/10.1007/s00259-008-1051-3.
[38] Meyer S, Mueller K, Stuke K, Bisenius S, Diehl-Schmid J, Jessen F, et al. Predicting behavioral variant frontotemporal dementia with pattern classification in multi-center structural MRI data. Neuroimage Clin 2017;14:656–62. https://doi.org/10.1016/j.nicl.2017.02.001.
[39] Bron EE, Smits M, Papma JM, Steketee RME, Meijboom R, de Groot M, et al. Multiparametric computer-aided differential diagnosis of Alzheimer's disease and frontotemporal dementia using structural and advanced MRI. Eur Radiol 2016. https://doi.org/10.1007/s00330-016-4691-x.
[40] Office for National Statistics (ONS). Population estimates for UK, England and Wales, Scotland and Northern Ireland. 2018. Available at: https://data.gov.uk/dataset/849f9984-dfe4-46a5-8162-c5dee3f19ea4/population-estimates-for-uk-england-and-wales-scotland-and-northern-ireland (accessed November 19, 2018).
[41] Brzezicki MA, Kobetić MD, Neumann S. Frideswide – an artificial intelligence deep learning algorithm for audits and quality improvement in the neurosurgical practice. Int J Surg 2017;43:56–7. https://doi.org/10.1016/j.ijsu.2017.05.038.
[42] Chen JH, Asch SM. Machine learning and prediction in medicine — beyond the peak of inflated expectations. N Engl J Med 2017;376:2507–9. https://doi.org/10.1056/NEJMp1702071.
[43] Russell S, Norvig P. Artificial intelligence: a modern approach, global edition. 3 edition Boston Columbus: Pearson; 2016. ISBN-10: 1292153962.
[44] Hibbard B. Model-based utility functions. J Artif Gen Intell 2012;3:1–24. https://doi.org/10.2478/v10229-011-0013-5.
[45] Webb G, Yu X, AI. Advances in artificial intelligence. 17th Australian Joint Conference on Artificial Intelligence, Cairns, Australia, December 4-6, 2004, Proceedings. Springer; 2004. ISBN 978-3-540-30549-1 URL: https://www.springer.com/la/book/9783540240594.
[46] Pijnenburg YAL, Mulder JL, van Swieten JC, Uitdehaag BMJ, Stevens M, Scheltens P, et al. Diagnostic accuracy of consensus diagnostic criteria for frontotemporal dementia in a memory clinic population. Dement Geriatr Cogn Disord 2008;25:157–64. https://doi.org/10.1159/000112852.
[47] Mendez MF, Shapira JS, McMurtray A, Licht E, Miller BL. Accuracy of the clinical evaluation for frontotemporal dementia. Arch Neurol 2007;64:830–5. https://doi.org/10.1001/archneur.64.6.830.
[48] Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D. Concrete problems in AI safety. 2016. ArXiv160606565 Cs. URL: https://arxiv.org/abs/1606.06565.