BMC Medical Research Methodology

RESEARCH

Open Access

Evaluation of data imputation strategies in complex, deeply-phenotyped data sets: the case of the EU-AIMS Longitudinal European Autism Project

A. Llera[1,2,3*], M. Brammer[4], B. Oakley[5], J. Tillmann[6], M. Zabihi[1,2], J. S. Amelink[1,7], T. Mei[1,2], T. Charman[8], C. Ecker[4,9], F. Dell'Acqua[4], T. Banaschewski[10], C. Moessnang[9,11], S. Baron-Cohen[12], R. Holt[12], S. Durston[13], D. Murphy[4,5], E. Loth[4,5], J. K. Buitelaar[1,2,14], D. L. Floris[1,2,15†] and C. F. Beckmann[1,2,16†]

Abstract

An increasing number of large-scale multi-modal research initiatives has been conducted in the typically developing population, *e.g.* Dev. Cogn. Neur. 32:43-54, 2018; PLoS Med. 12(3):e1001779, 2015; Elam and Van Essen, Enc. Comp. Neur., 2013, as well as in psychiatric cohorts, e.g. Trans. Psych. 10(1):100, 2020; Mol. Psych. 19:659–667, 2014; Mol. Aut. 8:24, 2017; Eur. Child and Adol. Psych. 24(3):265–281, 2015. Missing data is a common problem in such datasets due to the difficulty of assessing multiple measures on a large number of participants. The consequences of missing data accumulate when researchers aim to integrate relationships across multiple measures. Here we aim to evaluate different imputation strategies to fill in missing values in clinical data from a large (total $N=764$) and deeply phenotyped (i.e. range of clinical and cognitive instruments administered) sample of $N=453$ autistic individuals and $N=311$ control individuals recruited as part of the EU-AIMS Longitudinal European Autism Project (LEAP) consortium. In particular, we consider a total of 160 clinical measures divided in 15 overlapping subsets of participants. We use two simple but common univariate strategies—mean and median imputation—as well as a Round Robin regression approach involving four independent multivariate regression models including Bayesian Ridge regression, as well as several non-linear models: Decision Trees (Extra Trees., and Nearest Neighbours regression. We evaluate the models using the traditional mean square error towards removed available data, and also consider the Kullback–Leibler divergence between the observed and the imputed distributions. We show that all of the multivariate approaches tested provide a substantial improvement compared to typical univariate approaches. Further, our analyses reveal that across all 15 data-subsets tested, an Extra Trees regression approach provided the best global results. This not only allows the selection of a unique model to impute missing data for the LEAP project and delivers a fixed set of imputed clinical data to be used by researchers working with the LEAP dataset in the future, but provides more general guidelines for data imputation in large scale epidemiological studies.

Keywords: Imputation, Clinical data, Multivariate, Machine learning

Introduction

In clinical settings, a broad array of data using questionnaires, observational methods or interviews, and behavioural assessments is acquired that involve a number of individuals (n) and a number of clinical variables (p).

†D. L. Floris and C. F. Beckmann shared last authorship.

*Correspondence: a.llera@donders.ru.nl

[1] Donders Institute for Brain, Cognition and Behaviour, Centre for Cognitive Neuroimaging, Nijmegen, The Netherlands
Full list of author information is available at the end of the article

Missing data is a general problem in data analyses [1–6] since most algorithms cannot directly handle the presence of missing values. Although there exist models able to handle missing observations, these are scarce, strongly tailored for specific analyses and consequently their use is limited and not a standard procedure [7–9]. Instead, the usual way researchers proceed in such cases is to reduce the sample size (n) by removing individuals with missing data variables, i.e. Available Case Analyses [10], resulting in a decrease of statistical power for any further analyses [11]. This problem becomes most notable when performing multi-modal analyses involving multiple variables [12, 13], for example classification or clustering, since the number of individuals available in any such analyses will be limited by the simultaneous availability of several clinical measures, reducing the sample size even further. A reduced sample size has a direct effect on the statistical power resulting in reduced sensitivity to and specificity of findings as well as limiting the degree to which sample heterogeneity can be investigated. This is problematic especially in cases where a small effect is usually expected, as it is the case for example in computational psychiatry. At the same time, an increased sample size will also provide more confidence in the observed patterns and increases reproducibility. Other important issues when excluding participants due to one or more missing values are both the associated 'economic cost' in the sense of not utilising all the (research) resources invested in the study, and the ethical issue of "human cost", i.e. high time investment on the part of the experimenter and the participants during data collection. Further, data loss can have an even bigger impact on analyses where one wants to study the relationship between different data modalities, such as clinical/behavioural variables and neuroimaging or genetic data [14, 15]. Basically, missing clinical measures reduce the full imaging/genetic sample resulting in a significant loss of statistical power, and a dramatic under-utilisation of investment on the part of funders, researchers and research participants. This is particularly a problem in the case of big-data consortia where a wide range of expensive data collections are performed [16–25] . An alternative approach to deal with missing data values is *data imputation* [30]. This approach substitutes missing values by applying a statistical estimation of their values, and consequently avoids reducing the sample size and prevents associated loss issues. A very common and simple strategy for imputation of behavioural or clinical data is substituting individual missing values by the mean or the median of the observed sample values of the respective variable. Even though this approach allows one to retain the original sample size, it does not improve the statistical power of consequent analyses, the reason being that the number of independent clinical observations (the 'true' degrees of freedom for a particular measure) remains fixed. Furthermore, such simplistic imputation strategies are not well suited when heterogeneity can be expected in the clinical group, e.g. when the distribution of observed values is not unimodal. A more advanced strategy which circumvents this shortcoming of mean/median imputations, and is thus able to increase the amount of independent observations, is based on multivariate regression models [31]. These use *all* clinical variables to obtain expectations over the values at each missing value per variable [32]. Such an approach typically uses a Round-Robin [33, 34] scheduled regression where missing values expectations are iteratively updated through all variables until convergence of the distribution of the missing values is reached. This procedure can be found in the literature under different notations as imputation by means of chained equations [34], sequential regression imputation [35] or more generally, fully conditional specification [36]. In such approaches, every missing value expectation for a given variable is different for different participants since it is based on the observations and expectations of all variables for each participant independently. Consequently, this approach increases the number of independent observations with respect to the simpler univariate imputation approaches. Obviously, Round-Robin multivariate regression strategy results are dependent on the regression model chosen, and in fact, this choice is the biggest difference between the most common imputation packages used in practice. For example, some common packages use parametric regression procedures [34], whereas others use non-parametric regression models [37], all cases embedded in a Round-Regression scheduling process. Such models can in addition be evaluated with several random initializations, i.e. multiple imputation [38], to obtain statistics reflecting also the uncertainty over the estimated parameters.

In this work, we use behavioural/clinical data from the European Autism Interventions Multicenter Study (EU-AIMS), Longitudinal European Autism Project (LEAP) consortium – the largest, international multi-centre initiative dedicated to identifying biomarkers in Autism Spectrum Disorder (henceforth 'autism'). To study autism at the neurobiological and genetic level, data were collected from a population of individuals with an autism diagnosis as well as from typically developing (TD) individuals between 6–30 years of age. The sample is deeply phenotyped with an extended battery of behavioural, cognitive and clinical assessments alongside a wide range of quantitative measurements such as electroencephalogram, structural and functional magnetic resonance imaging, biochemical markers and genomics [17]. In the LEAP sample in particular, and in most large-scale

imaging consortia in general, missing behavioural and clinical data has a large impact due to the extensive and expensive battery of imaging and genetic data acquired. Consequently, clinical data imputation has shown itself necessary to fully exploit the potential of such a rich and valuable dataset. The need becomes even more evident in the context of longitudinal study designs such as LEAP, where missing behavioural and clinical data at one timepoint poses additional challenges for meaningful longitudinal analyses. The aim of the present work is to perform a systematic and extensive evaluation of different imputation models to be able to provide a state-of-the-art imputation procedure for the EU-AIMS LEAP cohort in particular and provide a unique set of imputed data to use for all researchers involved in LEAP. Consequently, our present work aims to avoid biases resulting from different researchers using different models to impute clinical data for their future individual analyses when relating for example brain or genetics data to clinical measures. Since the evaluation of such models is not trivial, we develop quantitative measures to assess the quality of the imputation.

Methods

The dataset

EU-AIMS LEAP is the to-date largest multi-centre, multi-disciplinary observational study on biomarkers for autism involving a large sample of 764 individuals including 453 autistic children, adolescents and adults and 311 TD individuals (or with mild intellectual disability [ID] without autism) between the ages of 6 and 30 years. Each individual is comprehensively characterised at multiple levels including their clinical profile, cognition, brain structure and function, biochemistry, environmental factors and genomics. This study utilises an 'accelerated longitudinal design', comprising four cohorts defined by age and ability level: Children with either autism or typical development aged 6–11 years and intelligence quotient (IQ) in the typical range, adults with either autism or TD aged 12–17 years and IQ in the typical range, young adults with either autism and TD aged 18–30 years and IQ in the typical range, and adolescents and adults with mild intellectual disability with/without autism aged 12–30 years [17, 39]. The study involves a comprehensive approach to deep phenotyping. Due to differences in age and ability level, measures were divided by experimental design into core measures that were assessed in all participants, and measures that were selectively administered in some schedules which were appropriate for adolescents and/ or adults with higher cognitive function but not for children or those with mild ID. This includes questionnaire measures, such that parents were used as informants in all schedules (except for typically developing adults, where parents were not available to participate in the study) while self-report questionnaires were only used in adolescents and adults. We also aimed to reduce the testing burden of experimental tests (e.g., magnetic resonance imaging [MRI] acquisition times) for children and young people with ID. The full protocol includes a) demographics, such as education of caregiver and parental household income or medical history, b) observational measures of autistic features (e.g., Autism Diagnostic Observation Schedule [ADOS] [40]), c) parent-based interviews (e.g., Autism Diagnostic Interview [ADI-R] [41], Vineland Adaptive Behaviour Scale [VABS-II] [42]), d) parent- and self-reported questionnaires of the core autism phenotype (e.g., Social Responsiveness Scale [SRS-2] [43]; Repetitive Behavior Scale [RBS-R] [44]; Short Sensory Profile [SSP] [45]), associated features (e.g., Sleep Habit Questionnaire [46], Empathy Quotient [47–49], Child Health and Illness Profile [50] and measures of commonly co-occurring conditions (e.g., Attention-Deficit/Hyperactivity Disorder [ADHD]: DSM-5 ADHD rating scale; Strenghts and Difficulties Questionaire [SDQ] [51]; Development and Well-Being Assessment [DAWBA] [52], anxiety: Beck Anxiety Inventory [53], depression: Beck Depression Inventory [54]). We deliberately included several questionnaires that overlapped in their construct content, e.g. assessing core features of autism, to validate them externally. This means that high correlations between some measures were expected. The protocol further includes e) cognitive assessments, including e.g., Intellectual functioning (IQ): Wechsler Intelligence Scale for Children (WISC) [55], Wechsler Adult Intelligence Scale (WAIS) [55] handedness: Edinburgh Handedness Inventory [56], social cognition, (e.g., theory of mind: animated shapes task [57]; false belief task [58]); executive function Spatial Working Memory [59]. Some cognitive tests used behavioural response variables while others also acquired functional brain responses (e.g., using functional MRI [fMRI] Flanker task [60], Social and Non-Social Reward task [61], or electroencephalogram [EEG, e.g., mismatch negativity, face processing]). A detailed description of the clinical cohort and extended characterisation can be found in [17, 39]. In this paper we consider a set of 160 clinical measures in total, including 2 nominal binary variables that contain no missing values (diagnosis and sex), 42 continuous valued variables and 116 ordinal valued variables. A complete detailed list of all included measures in the analyses is provided as Supplementary Table 1 (ST1).

The 160 measures considered in this paper expand self and parent reported measures, and include a subset of measures acquired for all 764 participants, a subset acquired for all 453 individuals with autism, and several

other subsets of measures acquired uniquely for subsets of individuals defined by four different enrolment schedules (adults, adolescents, children or intellectual disability [ID]. This resulted in a total of 15 different subsets structured based on group (autism vs. TD), schedule and acquisition method. A summary of all these subsets of participants for which measures are present is summarized in Table 1, where a total of 15 different subsets of individuals and measures are defined. A summary of the number of variables (p), individuals (n), percentage of missing samples as well as the target group (i.e., diagnostic group and enrolment schedule) in which the measure was supposed to be acquired in the first place (i.e., green vs. not acquired in the group = red).

In Fig. 1, we show the correlation structure of all these variables, grouped by subsets as indicated by the horizontal and vertical black lines. We observe that some subsets do not share participants (white areas), and also that many measures are intercorrelated inside and across subsets, providing a primary motivation for multivariate imputation strategies. More detailed information about the variables included in each of these subsets can be found in Supplementary Table 1.

There are 28 core clinical/behavioural/demographic measures that include all 764 individuals (subset 1), and these measures include for example, age, sex, IQ or handedness. In subset 2, we observe that there are 8 measures comprising all the 453 autistic individuals which include ADOS and ADI. Subset 3 comprises 653 participants and includes all TD individuals along with all autistic children and adolescents; it includes 30 measures with some examples being repetitive behaviour or short sensory profile measures. Subset 4 excludes also TD adolescents from subset 3 and involves the Vineland Adaptive Functioning Scale [42]. Subset 5 includes TD and autistic individuals, but excludes individuals with ID; this includes a total of 653 individuals and 4 cognitive task measures involving Hariri [62] and theory of mind tasks [57, 63]. Subset 6 excludes all children from subset 5, resulting in a total of 478 individuals and 32 clinical measures as for example Flanker [60, 64] or Social Responsive Scale tests [65]. Subset 7 is also acquired for TD and autistic individuals but excludes adults and individuals with ID older than 18 years, including a total of 458 participants and 6 measures, such as Children Social Behavior Scale (CSBQ) [66, 67] and Child Health and Illness Profile (CHIP) [72, 73] questionnaires. Without need for further specification of the details for the remaining subsets, it is clear that the individuals included in any of these subsets, are also partially contained in other subsets, and the full picture is a complex organisation of participants and measures (based on diagnostic group, schedule and acquisition type). As a consequence of such a complex structure of clinical data gathering, one cannot use all variables for direct imputation of all the other ones since it would not be sensible to impute data that was not supposed to be acquired in a certain group at the first stage which would result in bias. For example, it would not be appropriate to impute ADI or ADOS measures in TD individuals, as in this study we did not attempt to acquire ADI and ADOS on the TD participants. It is important to note that these 15 subsets of clinical measures have very different properties. First, in terms of the ratio of observations to number of variables, n/p (see Table 1). As such, the performance of any regression model can be expected to be different on each subset, even in the hypothetical case of non-missing data. For completeness let's remember that a higher n/p ratio allows more robust and reliable learning [74, 75]. Second, higher percentage of missing values makes the estimation of the missing values harder.

In Fig. 2 we visualize some characteristics of the missing data itself, with each row presenting one of the 15 subsets. The left column illustrates the missing values themselves as blue dots, with participants represented in the x-axis and the number of variables included on that subset of the full data in the y-axis. For example, we can observe that subset 1 contains a few measures with no missing values (rows with no blue dot) which include diagnosis, age and sex. In general, for all subsets we can appreciate that white vertical lines show individuals with many variables acquired, while white horizontal lines index measures acquired for many individuals.

In the second and third columns, we color-coded the percentage of shared missing variables between each pair of individuals and the percentage of shared missing individuals between each pair of variables respectively. In these two columns, darker coloured areas index pairs of individuals or measures with many missing shared values respectively. Fourth and fifth columns present histograms of the number of individuals and variables missing respectively. The sixth column presents the correlation between the variables on each subset, where the non-diagonal images show the correlated structure on these measures which motivates the use of multivariate models to estimate their missing values also on each of the subsets independently.

Imputation strategies

For the remainder of this paper we denote by n the number of individuals, by p the number of variables, and by m the number of missing values, where $m = \sum_{j=1}^{p} m_j$ and m_j denotes the number of missing observations for the j^{th} variable. Consequently, we consider the imputation of a data matrix D of size n times p, where there are m missing values and we denote as D^* the imputed data

Table 1 All clinical data from the EU-AIMS LEAP consortium acquired at wave 1 is summarised as 15 different subsets as indicated in each row. The columns show the number of variables and participants included on each of these subsets as well as the percentage of missing data. Color-coded columns indicated the availability (green) or lack of data (red) as acquired for a subgroup of the participants as indicated in each column

Subset	Variables (p)	n	n/p	% missing	Groups	Schedules			
						Adult	Adolescents	Children	ID
1	28	764	30.6	16.1	ASD				
					TD				
2	8	453	56.6	12.5	ASD				
					TD				
3	30	653	21.8	27.9	ASD				
					TD				
4	4	560	140	17.9	ASD				
					TD				
5	4	653	163.3	32.8	ASD				
					TD				
6	32	478	14.9	36.7	ASD				
					TD				
7	6	458	76.3	17.7	ASD				<18 yo
					TD				<18 yo
8	1	201	201	31.3	ASD				>18 yo
					TD				>18 yo
9	2	235	117.5	19.2	ASD				<18 yo
					TD				<18 yo
10	14	255	18.2	37.1	ASD				
					TD				
11	6	223	37.2	26.7	ASD				
					TD				
12	14	175	43.8	19.1	ASD				
					TD				
13	8	111	13.9	41	ASD				
					TD				
14	2	57	28.5	44.7	ASD				>18 yo
					TD				>18 yo
15	1	334	334	35.9	ASD				
					TD				

Abbreviations: *ASD* Autism spectrum disorder, *TD* Typically developing individuals, *ID* Intellectual disability

matrix. We consider the use of six imputation strategies including two simple but common univariate strategies, mean and median imputation, as well as four multivariate regression models including a linear model, Bayesian Ridge (BR) regression [76], as well as several non-linear models, Decision Trees (DT) [77], Extra Trees (ET) [78] and Nearest Neighbours (NN) [71]. Table 2 provides an overview of these models. Since all discrete variables requiring imputation in this dataset are ordinal, and some can take a high number of possible values, we decided to

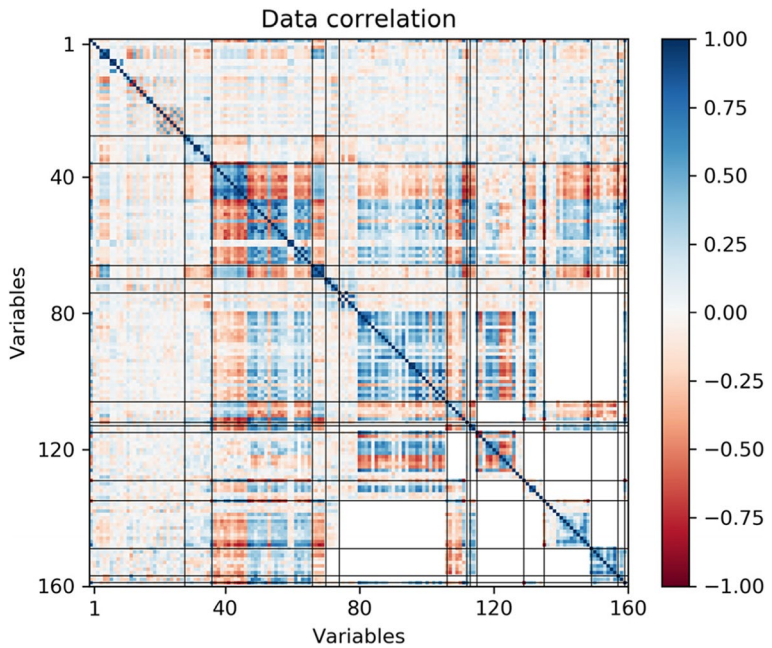

Fig. 1 Correlation structure of the 160 clinical measures. White areas correspond to subsets of measures with no shared participants

avoid using classification models for the ordinal variables and impute them all using regression models followed by rounding when needed [79].

The univariate imputation strategies substitute all the missing observations at each variable $j\epsilon\{1\dots p\}$ by some relevant summary statistics at the non-missing values at that variable, i.e. some statistics at the available entries at the j^{th} column of D. In particular here we consider the mean and median imputation strategies.

Such strategies are suboptimal from both a statistical and a clinical point of view; from a statistical point of view they ignore the correlation of the data shown in Figs. 1 and 2, and from a clinical point of view, since we know that autism, as many other neurodevelopmental and neuropsychiatric conditions, is clinically and etiologically heterogeneous, meaning that we already a priori assume that there are different relationships between clinical variables and underpinning mechanisms in potentially different subgroups.

These facts strongly motivate moving towards multivariate models for imputation. In the case of multivariate methods, since all variables are needed for imputation of each single variable missing values, we use a Round-Robin [33] regression approach, treating every variable as an output in turn. This approach requires defining an order for variable imputation. For simplicity, here we consider an ordering where variables are imputed in an ascending order of number of missing values. Initially, once the first variable of interest to be imputed is selected according to the chosen variable ordering, all other variables missing data values are set to its expected value using mean imputation, and the considered multivariate regression model is used to obtain an expectation of the missing values on the variable of interest. Then the next variable of interest is selected according to the ordering and the originally missing values are estimated as above. The process is repeated for all variables to close the first round of the Round-Robin iterative process and obtain estimations for all missing values that are consequently different from the initial mean imputation values assigned. The Round-Robin cycle is repeated as many times as needed, using at each round the estimated

(See figure on next page.)

Fig. 2 Each row presents information about one of the fifteen subsets. The first column (left) presents missing data as blue dots with individuals presented in the x-axis and number of clinical measures in the y-axis. The second and third columns present the percentage of shared missing variables per pair of individuals, and the percentage of missing individuals per pair of behavioural measures respectively, with darker colours coding an increased percentage. The fourth and fifth columns present histograms showing the number of individuals missing a number of variables, and the number of variables being missed by a number of individuals. The sixth column present the correlation structure inside each of the subsets i.e. diagonal subsquares of Fig. 1

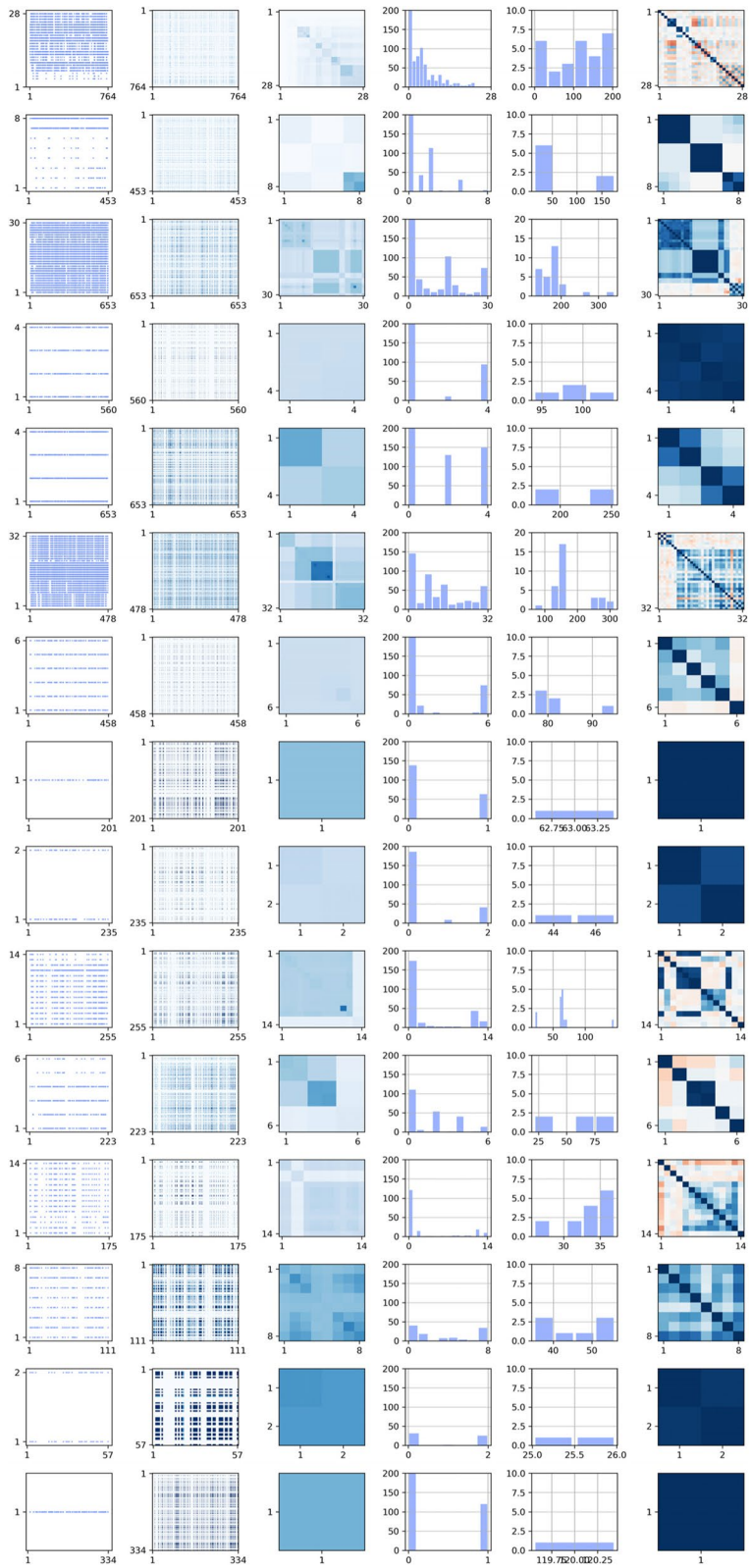

Fig. 2 (See legend on previous page.)

Table 2 Imputation strategies considered

		Imputation Strategy
Univariate		Mean
		Median
Multivariate Regression with Round-Robin schedule	Linear	Bayesian Ridge (BR)
	Non-linear	Decision Trees (DT)
		Extra Trees (ET)
		Nearest-Neighbours (NN)

missing values from the previous round, until all imputed values at all variables converge. Here we set to 100 the maximum number of Robin-Rounds to perform. All imputations were performed using publicly available tools [80].

Order of imputation

As shown in Table 1, the clinical data breaks down to a very complex organization of measures according to the population for which they are acquired, which can be summarized as 15 different subsets of data. Consequently, it would not be sensible to impute for certain individuals measures that were not intended to be acquired for them in the experiment design. However, imputation of each of the 15 subsets independently would be suboptimal since we observed correlations also across subsets in Fig. 1. Consequently, one needs to combine subsets to maximise the imputation power. To that end, we performed an exhaustive search to find the optimal order of imputation of each of these subsets, while for imputation of a target subset we used any previously imputed subsets, as long as the target population is contained in the previously imputed subsets.

The process starts with the imputation of subset 1 in isolation, since all participants were planned to be measured with respect to these 28 variables. It is important to mention that from subset 1 we removed the clinical measure 'diagnosis' so as to not bias the imputation towards the diagnosis label and avoid producing a bias effect in any posterior study on these imputed data. Our brute force optimization showed that the next subset to impute should be subset number 3, which is acquired for all participants with the only exception of autistic adults; for imputation of subset 3 we used the imputed values of subset 1, restricted to the individuals in subset 3, in addition to the variables on subset 3. After we proceeded to subset 4 and then to subset 2. In Table 3 we provide the structure of the ordering performed to maximize the power of all the imputation process, where an asterisk denotes an imputed file. The fourth column indicates the already imputed files that are considered for imputation of each input file.

Table 3 Order followed for imputation of the subsets. The last column shows the imputed subsets used for imputation of each subset indicated in the second column

order	input	output	Conditioned to
1st	Subset 1	Subset 1*	None
2nd	Subset 3	Subset 3*	Subset 1*
3rd	Subset 4	Subset 4*	Subsets 1*, 3*
4th	Subset 2	Subset 2*	Subsets 1*,3*,4*
5th	Subset 5	Subset 5*	Subset 1*
6th	Subset 6	Subset 6*	Subset 1*,5*
7th	Subset 7	Subset 7*	Subsets 1*, 3*
8th	Subset 15	Subset 15*	Subsets 1*,3*
9th	Subset 8	Subset 8*	Subsets 1*,3*,4*
10th	Subset 9	Subset 9*	Subsets 1*,3*,4*,7*
11th	Subset 10	Subset 10*	Subsets 1*,5*,6*
12th	Subset 11	Subset 11*	Subsets1*,3*,5*,6*,7*,15*
13th	Subset 12	Subset 12*	Subsets 1*,3*,4*,5*,7*,9*
14th	Subset 13	Subset 13*	Subsets1*,3*,4*,15*
15th	Subset 14	Subset 14*	Subsets1*,3*,4*,8*,13*,15*

Note that as a result of such experimental design, when considering all 160 measures in our sample together, there is a systematic relationship between the propensity of missing values at certain variables and the observed data. For example, some measures (subset 10) are acquired for adults only, while age is also an available variable. Consequently, when considering all 160 measures together, missing data at some variables is most probably *missing at random* (MAR) [81]. Although one cannot distinguish between MAR and *missing not at random* (MNAR) [81] without a follow up intervention in the dataset, field expertise and careful data gathering, suggests the absence of a MNAR structure in the variables of our dataset. Further, when considering the imputation of each subset independently, or when following the order of imputation for the different subsets we introduced here, each subset is imputed using only subjects of corresponding diagnosis group, age or IQ range, making the missing data on each subset most probably Missing Completely At Random (MCAR) [81]. Although there exist tools to get insights into whether data is MCAR or MAR [26, 27], it has been shown that in both cases unbiased estimations can be obtained using iterative imputation schemes [28].

Evaluation

There is need for a strict validation of the imputation results since the imputation choice can have a strong bias effect on the clinical-brain/genetics associations which needs to be minimized. To quantify the quality of each imputation model we use two different measures.

1) We first compute the quality of the imputation using a leave-one-observation-out cross-validation approach. More exactly, for each imputation model, we perform $(nxp) - m$ imputation problems, where at each of the problems we add an extra missing value to the original problem, let's say at location (i,j), resulting in a data matrix to be imputed with m+1 missing values. This means that $D_{i,j}$ is an originally observed value that has been artificially removed in a fold of the cross-validation loop to be able to evaluate the imputation error at location (i,j) by comparison with respect to the imputation value obtained at that location, $D^*_{i,j}$. For clarity of notation we denote the variable indexes in D as $j\epsilon\{1,\ldots,p\}$, and the originally available observations indexes at the j-th variable in D as $i\epsilon\{k_{j,1},\ldots,k_{j,n-m_j}\}$. After performing the imputation using any selected imputation model to obtain an imputed data matrix D^*, we compute the total error at the removed value D_{ij} as

$$E(i,j) = \sqrt{\left(D_{ij} - D^*_{ij}\right)^2}$$

To have a measure of error considering the scale of each variable independently, we define a relative error (RE) measure by dividing the observed and imputed values in E by the mean of the observed values at D, per each variable j independently. That is

$$RE(i,j) = \sqrt{\left(\frac{D_{ij}}{\mu_j} - \frac{D^*_{ij}}{\mu_j}\right)^2} = \sqrt{\frac{(D_{ij} - D^*_{ij})^2}{\mu_j^2}} = \frac{\sqrt{(D_{ij} - D^*_{ij})^2}}{|\mu_j|} = \frac{E(i,j)}{|\mu_j|}$$

where $\mu_j = \frac{1}{n-m_j}\sum_{k\epsilon O_j} D_{kj}$.

Consequently $RE(i,j)$ is simply a scaled version of $E(i,j)$ that relates to the size of the error with respect to the size of the variable values, and assigns a value of 0 in the case of no estimation error and a value of 1 when the error (E) is of the size of the mean observed value at that variable. Such representation facilitates the comparison of values on *RE* across variables taking values at different scales. Finally, to summarize *RE* per variable we take its mean value across the observations at that variable and we denote it as

$$MRE(j) = \frac{1}{n-m_j}\sum_{k\epsilon O_j} RE(k,j), \forall j\epsilon\{1,\ldots,p\} \quad (1)$$

2) We use the Kullback–Leibler (KL) divergence [75] to measure the overall effect of data imputation to the distribution of values. The KL divergence assigns a value of zero to identical distributions, and increasing values to distributions that deviate from each other. We perform the imputation of the original data matrix D and compute, at each variable independently, the KL divergence between the initially observed distribution and the distribution of estimated values at the missing participants. More precisely,

$$KL(p_j||q_j) = \sum_x p_j(x)\log\left(\frac{p_j(x)}{q_j(x)}\right), \forall j\epsilon\{1,\ldots,p\} \quad (2)$$

where $p_j(x)$ is the distribution of the observed values at the j^{th} variable and $q_j(x)$ the distribution of the imputed missing values at that same variable [74].

It is to note that the amount of missing values for a particular measure, is, to a certain degree, induced by the experimental design. The reason is that measures were acquired in a defined order of relevance because it was expected that several participants might not complete all questionnaires. Consequently, by experimental design, there are more subjects missing specific sets of variables which might result in a bias in the cross-validation MRE at these variables. This bias could occur since artificially removed values might be easier to estimate than actually missing values (because during the iterative imputation one may not rely on expected values from other variables but rather on real observations). Consequently, the MRE might be underestimated in the cross-validation setting and not represent the true generalization error in truly missing values. This motivates the introduction of the second measure of error, the KL divergence, that will penalize models providing distributions at the missing values that deviate from the observed distribution.

Although each of these performance measures is informative for each variable, they cannot simply be combined since they quantify mismatch at different scales. However, we can build a two-dimensional error function by considering the MRE and KL values per variable relative to some reference model. Consequently, to be able to consider simultaneously the MRE and the KL measures of error, and to be able to pull many variables together to draw any conclusion, we define as a reference model the mean imputation model, and divide for each variable, the MRE and the KL measures at each model by the MRE and KL values obtained by the mean imputation model. In this way, we obtain *MRE and KL measures relative to the mean imputation*, assigning for each variable the mean imputation performance to the plane point (1,1), and all other performances can be pulled together as they represent a relative improvement with respect to the mean imputation. Consequently, for a given variable and a fixed imputation model, we consider the robenious norm of such two-dimensional 'error vector', i.e. the

square root of the sum of absolute squared values in the error vector [29], as a global measure of error that combines both MRE and KL.

Results

Following the ordering of the 15 subsets of clinical measures indicated in Table 3, we proceeded to the imputation of the missing values in the clinical dataset from EU-AIMS LEAP. As illustrated in section "Methods: The dataset", each of these data matrices present different challenges to perform their imputation, with for example subset 6 being more challenging than subset 2, since the subset has a smaller n/p ratio and has many more missing values (see Table 1). Consequently, these 15 subsets serve as an interesting test bed to study the robustness of the different algorithms in general and not uniquely for this dataset, since we can check the performance in the harder problems in relation to the simpler ones.

Figure 3 shows the MRE and KL plane relative to the mean imputation for each subset (subplots), with each

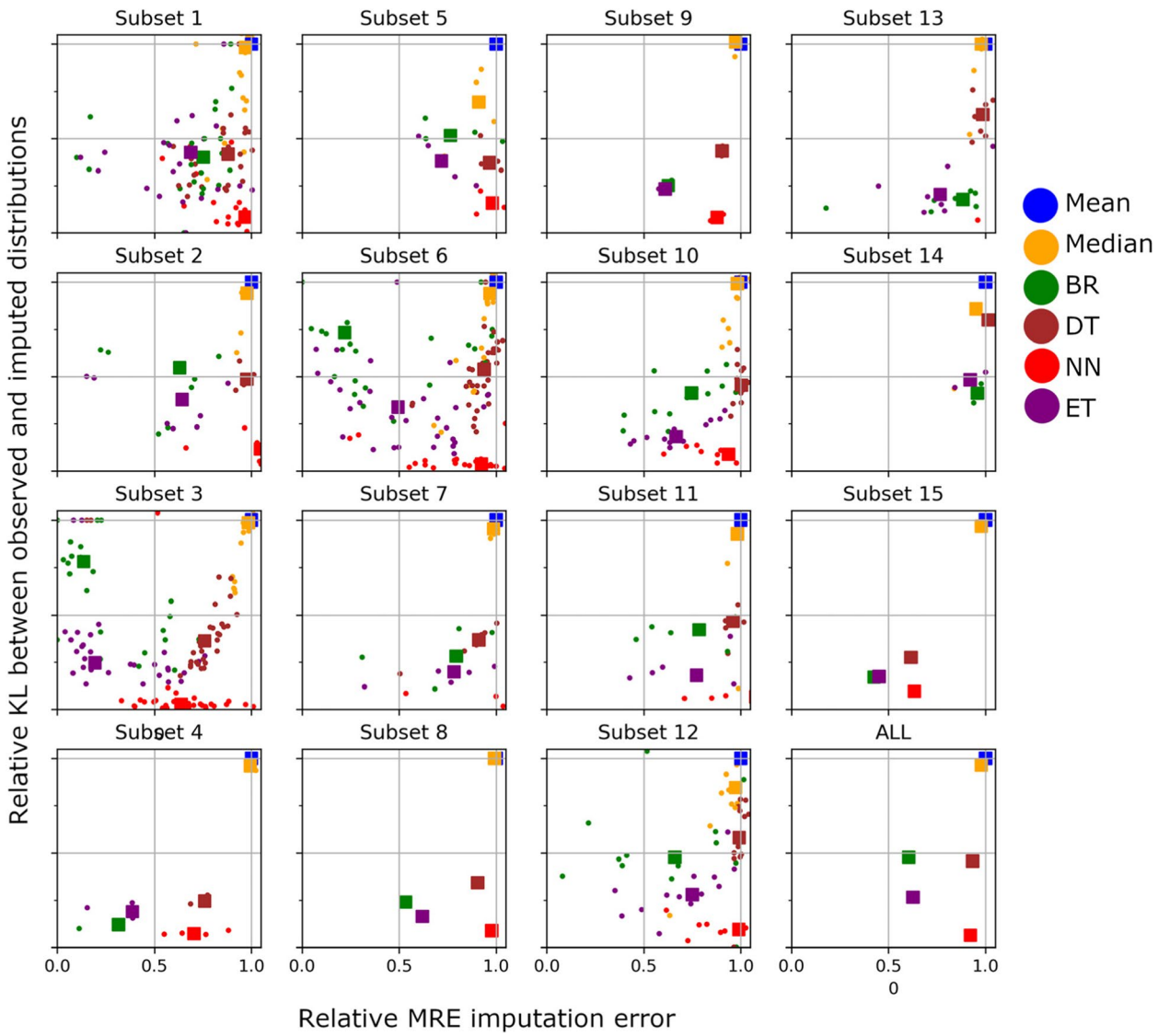

Fig. 3 Visualization of the imputation performance at the clinical measures acquired at each of the subsets. Each subfigure presents the performance for each clinical measure in the subset as dots, and for the 6 imputation models considered (color coded). The colored squares show the mean across measures per model. For each subset, the x-axis shows the mean imputation error (MRE) relative to the mean imputation model, and the y-axis the KL-divergence between the distribution at the availabel (observed) data and the imputed data at the missing values, again relative to the mean imputation model. Color coding in the legend: blue and yellow represent the univariate models, mean and median imputation respectively; green represents a multivariate linear Bayesian Ridge regression model (BR). The remaining colors encode multivariate non-linear models, with brown encoding decision trees (DT), red encoding nearest neigbours (NN), brown and purple extra tree regressors (ET)

dot representing one clinical variable in that subset, the different imputation models being color-coded and the colored squares representing the mean of the values for a given model in that subset. Further, the bottom right figure shows the mean performance of each model pulled across all measures of all 15 subsets. Recap for interpretation that models that are lower with respect to the y-axis perform better with respect to the KL divergence, while models that are plotted more to the left with respect to the x-axis perform better with respect to the MRE measure. Globally, models closer to (0,0) perform better. We first observe that in general the mean and median imputation perform much worse than all other models with respect to the MRE and also to the KL divergences i.e. blue and yellow dots show highest error. This is clear evidence for superior performance of multivariate models for such clinical measures imputation. With respect to the multivariate models we appreciate that NN performs well with respect to the KL, which makes sense since by looking at some of the closest neighbours it is allowed to sample the full space and get a distribution closer to the initially observed one. However, NN fails to provide a robust improvement with respect to the MRE, and in some subsets is even worse than the mean imputation (red squares not appearing in figure, for example for subset 13). From the remaining three models, we observe that Extra Trees Regressor (purple) and Bayesian Ridge Regression (green) outperform Decision Trees (brown). Although both Extra Trees and Bayesian Ridge provide an impresive improvement with respect to the mean imputation in terms of MRE (~40% reduction of error), Extra Trees provides a bigger improvement with respect to the KL divergence (~75 vs ~55% reduction of KL). Another interesting observation is that the imputation of all subsets provide a similar pattern of organization of the models performances, showing the robustness of the models performances across all subsets. This is an interesting finding given the huge differences in the n/p ratios as well as in the number of missing observations on each subset (Table 1). This representation confirms that the median imputation provides a similar performance to the mean imputation and they are the less accurate from the considered models. It further shows that BR provides in general a very high relative MRE improvement, but a lower relative KL improvement than the other multivariate models. It further highlights that the Extra Tree regressor is the model performing best in expectation. In fact, to compare the best two models, a paired t-test between the norms of the 2-dimensional errors in relative KL vs MRE plane of the ET and the BR models showed a significantly reduced error in favor of the ET model ($t = 4{,}01$, p-value $< 9 \times 10^{-5}$).

Discussion

We performed a comprehensive analysis and evaluation of six different imputation methods to compare the weaknesses and strengths of different methodologies to perform imputation of clinical variables. To that end we used 15 different subsets of clinical variables from the EU-AIMS LEAP dataset that have considerable differences in terms of ratio between number of variables and number of observations (n/p) as well as in terms of percentage of missing data values. We used standard univariate imputation techniques, i.e. mean and median imputation, as well as several multivariate regression models, i.e. Bayesian Ridge, Random Forest, Extra Trees, Decision trees. All the multivariate models were involved in a Round-Robin iterative scheduling till convergence of all missing values estimations. We evaluated the imputation using two different error measures, computing the error at the originally observed data using a leave-one-observation-out cross-validation approach, and also by computing the KL-divergence between the observation distributions and the imputed value distributions at each variable independently. To be able to compare the results of all models we scaled both error measures with respect to the mean imputation performances to obtain a measure of improvement with respect to the simplest mean imputation model. Even though the considered subsets had very different characteristics, the expected improvement with respect to the simpler mean imputation resembled in both cases a very similar pattern showing that the models performed in a similar fashion at the simplest as well as the hardest/most complex scenarios. In particular we observed that Extra Tree Regression was likely to be the best model for imputation of this dataset. All models were initially independently evaluated using grid search in a set of model parameters and the solution with the best set of parameters per model was selected and presented in this paper. In particular, for the Extra Tree Regression model we found that a model with 10 trees provided the best solution. Note that the Round-Robin regression approach is also implemented in the R-package for imputation 'Multiple Imputation by chained equations' (MICE) [34] and in fact, the python package we used here for imputation [80] is inspired in MICE. A particularity of MICE is that it models categorical variables using logistic or multinomial regression and continuous variables using linear regression [68]. As such MICE has more flexibility than the presented Bayesian Ridge Regression model, since it is tailored to model specifically categorical variables. However, the Tree based methods we considered are also able to capture such categorical structure from the data, and also handle multimodal distributions or capture non-linearities between

all the variables that might be hard to model using MICE, or require strong modelling and data domain specific knowledge. This has been empirically shown in [69] where it was found that although the difference between tree based methods and parametric MICE is not big, tree based methods outperformed the parametric models. Note that handling multimodal distributions is necessary where high heterogeneity is observed and consequently of outmost importance in the autism research where stratification based on clinical and imaging data is expected. One added particularity of MICE is that it runs the imputation problem many times with different initializations, returning finally the average of these imputations as final value. The most interesting of this approach is that it provides the standard deviation over the imputed values which serves as a measure of reliability in the imputation. Note that our extensive analyses also perform a validation that allows to get a measure of the quality of the imputation at each variable as given by the MRE and the KL divergences. In fact, the MRE evaluation performed is embedded in a cross-validation setting, where at each fold a different initialization is used. Since the error reported is the average of all the different folds, to a certain extent, it resembles the multiple imputation average scenario. However, we also considered a multiple imputation scenario for the best of our models, the Extra Tree Regressor. As suggested [80] we did not change the mean imputation as initialization but we rather used 100 different seeds to initially randomly build the regression trees. The results showed a standard deviation of order 10^{-3} at all the variables, showing that the estimation obtained using Extra Tree Regressors is extremely robust. Another similarity between the models employed in this work and well known models commonly used come from Random Forest regression embedded on Round-Robin scheduling being equivalent to another common package, missForest [37]. Although we did not include the full evaluation of Random Forest in this work, we performed several analyses during the preliminary preparation of this work and we observed that it would not improve ET or BR, its convergence was less satisfying, and the computational cost was orders of magnitude bigger. Our choice for python software [80] is driven by the flexibility of the packages to implement several regression models within the same framework, making the comparison between different models simpler and less error prone. We believe that the choice of model, and not of software, is critical for the quality of the imputation.

The Round-Robin scheduling procedure requires defining a variable ordering for imputation, and although here we report results using an increasing number of missing observations for variable ordering, results using a decreasing order did show similar results, both in terms of squared error and in terms of KL divergences between the observed and the imputed distributions at most variables, and for most models. Also, the patterns of models performances were identical. In conclusion, we systematically searched the best practice scenario for imputation of the clinical variables in this sample and found that Extra Trees Regressor was in expectation the best model. Given the different characteristics of the 15 data samples we consider that these results might also extrapolate to different datasets. As a result of this analyses we deliver the tools for imputation comparison we developed at https://github.com/allera/Imputation, and deliver imputed data to the EU-AIMS LEAP consortium; the neglectable standard deviation of the estimators obtained in the validation of the multiple imputation scenario using Extra Trees Regressors allows providing a unique dataset of imputed values.

A natural question arising is whether we can synthetically generate other missing measurements from such big data consortiums as for example structural brain images. The presented models are useful in their own for different types of vector data, however, models implementing spatial constraints should be more appropriate to interpolate data where a clear non-isotropic spatially smooth 3d distribution is expected. Ongoing research focuses on the imputation of missing structural MRI images, using existing structural MRI images and behavioural readouts, e.g. age, sex, weight. To that end we are considering extended convolutional neural networks [70] and we expect to be able to, for example, generate synthetic T1w images with smaller brain volume for younger participants. Once more, the quality of this approach can be validated by removing participants one at a time and checking the quality of the recovered image. Even more, given the relationship between structural features and functional features extracted from fMRI [14], we also aim to predict expected functional features based on structural and behavioural readouts, also using spatial convolution models. Such results are expected to follow up this work.

Abbreviations

EU-AIMS: European union autism interventions multicenter study; LEAP: Longitudinal European Autism Project; TD: Typical development; ASD: Autism spectrum disorder; BR: Bayesian Ridge regression; DT: Decision Tree; ET: Extra Trees; NN: Nearest Neighbours; MICE: Multiple Imputation by chained equations; KL: Kullback Leibler divergence; MRE: Mean root square error; SRS: Social Responsiveness Scale; RBS: Repetitive Behavior Scale; SSP: Short Sensory Profile; SDQ: Strenghts and Difficulties Questionaire; ADOS: Autism Diagnostic Observation Schedule; ADI: Autism Diagnostic Interview; DAWDA: Development and Well-Being Assessment; WISC: Wechsler Intelligence Scale for Children; fMRI: Functional MRI; EEG: Electroencephalogram.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-022-01656-z.

Additional file 1.

Acknowledgements
We thank all participants and their families for participating in this study. We also gratefully acknowledge the contributions of all members of the EU-AIMS LEAP group: Jumana Ahmad, Sara Ambrosino, Bonnie Auyeung, , Sarah Baumeister, Sven Bölte, Thomas Bourgeron, Carsten Bours, Daniel Brandeis, Claudia Brogna, Yvette de Bruijn, Bhismadev Chakrabarti, Ineke Cornelissen, Daisy Crawley, Guillaume Dumas, Jessica Faulkner, Vincent Frouin, Pilar Garcés, David Goyard, Lindsay Ham, Hannah Hayward, Joerg Hipp, Mark H. Johnson, Emily J.H. Jones, Prantik Kundu, Meng-Chuan Lai, Xavier Liogier D'ardhuy, Michael V. Lombardo, David J. Lythgoe, René Mandl, Andre Marquand, Luke Mason, Maarten Mennes, Andreas Meyer-Lindenberg, Nico Mueller, Laurence O'Dwyer, Marianne Oldehinkel, Bob Oranje, Gahan Pandina, Antonio M. Persico, Barbara Ruggeri, Amber Ruigrok, Jessica Sabet, Roberto Sacco, Antonia San José Cáceres, Emily Simonoff, Will Spooren, Roberto Toro, Heike Tost, Jack Waldman, Steve C.R. Williams, Caroline Wooldridge, and Marcel P. Zwiers.

Authors' contributions
AL, MB, EL, JB, DLF, and CFB have contributed to the study's conception. AL conducted the analyses, developed the evaluation techniques and generated figures. MZ, TM and JA optimized the toolbox code to deliver and edited latest manuscript versions and its revisions. DLF gathered and organized all data. AL and DLF generated tables and drafted the manuscript. BO provided all data. BO, JT, TC, CE, FDA, TB, CM, SB-C, RJH, SD, DM, EL, JKB and CFB have organized the EU-AIMS LEAP data and edited latest manuscript versions and its revisions; All authors have read and approved the final manuscript.

Funding
This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 115300 (for EU-AIMS) and No 777394 (for AIMS-2-TRIALS). This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA and AUTISM SPEAKS, Autistica, SFARI. This work has also been supported by the Horizon2020 programme CANDY Grant No. 847818). DLF is supported by funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101025785. This work was also supported by the Netherlands Organization for Scientific Research through VICI grant (Grant No. 17854 [to CFB]). The research leading to the presented work has received funding from the developing Human Connectome Project (dHCP) through a Synergy Grant by the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013), ERC Grant Agreement no. 319456. We also gratefully acknowledge funding from the Wellcome Collaborative Award (215573/Z/19/Z).
The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results. Any views expressed are those of the author(s) and not necessarily those of the funders.

Availability of data and materials
The data that support the findings of this study are available from the EU-AIMS Autism research in Europe but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available at the moment of submission but are available upon reasonable request to the corresponding author.

Declarations

Ethics approval and consent to participate
This study research, and all methods involved, were performed according to the regulations and guidelines defined by the Declaration of Helsinki. The experimental protocol was approved by the ethical licensing committee CMO Arnhem/Nijmegen under approval number CMO2014/288. Informed consent was obtained from all subjects and/or their legal guardians.

Consent for publication
Not applicable.

Competing interests
JKB has been a consultant to, advisory board member of, and a speaker for Takeda/Shire, Medice, Roche, and Servier. He is not an employee of any of these companies and not a stock shareholder of any of these companies. He has no other financial or material support, including expert testimony, patents, or royalties. CFB is director and shareholder in SBGneuro Ltd. TC has received consultancy from Roche and Servier and received book royalties from Guildford Press and Sage. DM has been a consultant to, and advisory board member, for Roche and Servier. He is not an employee of any of these companies, and not a stock shareholder of any of these companies. TB served in an advisory or consultancy role for Lundbeck, Medice, Neurim Pharmaceuticals, Oberberg GmbH, Shire, and Infectopharm. He received conference support or speaker's fee by Lilly, Medice, and Shire. He received royalities from Hogrefe, Kohlhammer, CIP Medien, Oxford University Press; the present work is unrelated to these relationships. JT is a current full-time employee of F. Hoffmann–La Roche Ltd. The other authors report no biomedical financial interests or potential conflicts of interest.

Author details
[1]Donders Institute for Brain, Cognition and Behaviour, Centre for Cognitive Neuroimaging, Nijmegen, The Netherlands. [2]Department of Cognitive Neuroscience, Radboud University Medical Centre, Nijmegen, The Netherlands. [3]LIS Data Solutions, Machine Learning Group, Santander, Spain. [4]Institute of Psychiatry, Psychology, and Neuroscience, Sackler Institute for Translational Neurodevelopment, King's College London, London, UK. [5]Department of Forensic and Neurodevelopmental Sciences, Institute of Psychiatry, Psychology, and Neuroscience, King's College London, London, UK. [6]Roche Pharma Research and Early Development, Neuroscience and Rare Diseases, Roche Innovation Center Basel, F. Hoffmann–La Roche Ltd., Basel, Switzerland. [7]Max Planck Institute for Psycholinguistics, Language & Genetics Department, Nijmegen, The Netherlands. [8]Department of Psychology, Institute of Psychiatry, Psychology, and Neuroscience, King's College London, London, UK. [9]Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, University Hospital Frankfurt Am Main, Goethe University, Frankfurt, Germany. [10]Child and Adolescent Psychiatry, Central Institute of Mental Health, University of Heidelberg, Mannheim, Germany. [11]Department of Applied Psychology, SRH University, Heidelberg, Germany. [12]Autism Research Centre, Department of Psychiatry, University of Cambridge, Cambridge, UK. [13]Department of Psychiatry, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, The Netherlands. [14]Karakter Child and Adolescent Psychiatry University Centre, Nijmegen, The Netherlands. [15]Methods of Plasticity Research, Department of Psychology, University of Zurich, Zurich, Switzerland. [16]Wellcome Centre for Integrative Neuroimaging - Centre for Functional MRI of the Brain (WIN FMRIB), University of Oxford, Oxford, UK.

Received: 10 February 2022 Accepted: 2 June 2022
Published online: 16 August 2022

References

1. Laird NM. "Missing data in longitudinal studies". Stat Med. 1988;7(1-2):05-315. https://doi.org/10.1002/sim.4780070131.
2. Schlomer GL, Bauman S, Card NA. Best practices for missing data management in counseling psychology. J Couns Psychol. 2010;57(1):1-10. https://doi.org/10.1037/a0018082.
3. Woodard JD, Shee A, Mude A. "A Spatial Econometric Approach to Designing and Rating Scalable Index Insurance in the Presence of Missing Data". The Geneva Papers on Risk and Insurance - Issues and Practice. 2016;41:259–79. https://doi.org/10.1057/gpp.2015.31.
4. Nogueira BM, Santos TRA, Zárate LE. "Comparison of classifiers efficiency on missing values recovering: Application in a marketing database with massive missing data". IEEE Symposium on Computational Intelligence and Data Mining. 2007 p. 66-72. https://doi.org/10.1109/CIDM.2007.368854.
5. Teegavarapu RSV, Tufail M, Ormsbee L. "Optimal functional forms for estimation of missing precipitation data". J Hydrol. 2009;374(1–2):106-15. https://doi.org/10.1016/j.jhydrol.2009.06.014.

6. Durham TJ, Libbrecht MW, Howbert JJ, Bilmes J, Noble WS. "PREDICTD PaRallelEpigenomics Data Imputation with Cloud-based Tensor Decomposition". Nat Commun. 2018;9:1402. https://doi.org/10.1038/s41467-018-03635-9.
7. Little RJA. "Regression with missing X's: a review". J Am Stat Assoc. 1992;87(420):1227-37. https://doi.org/10.1080/01621459.1992.10476282.
8. Chen T, Martin E, Montague G. "Robust probabilistic PCA with missing data and contribution analysis for outlier detection". Computational Statistics & Data Analysis. 2009;53(10):1, 3706-3716. https://doi.org/10.1016/j.csda.2009.03.014.
9. Von Hippel PT. "Regression with missing Ys: An improved strategy for analyzing multiply imputed data". Soc Methodol. 2007;37(1):83-117. https://doi.org/10.1111/j.1467-9531.2007.00180.x.
10. Pigott TD. "A review of methods for missing data". Educ Res Eval. 2001;7(4):353-83. https://doi.org/10.1076/edre.7.4.353.8937.
11. Nakagawa S, Freckleton RP. "Missing inaction: the dangers of ignoring missing data". Trends Ecology Evolution. 2008;23(11):592-6. https://doi.org/10.1016/j.tree.2008.06.014.
12. Vapnik VN. The Nature of Statistical Learning Theory. New York: Springer-Verlag; 1995.
13. Bishop CM. Pattern Recognition and Machine Learning. New York: Springer; 2006.
14. Llera A, Wolfers T, Mulders P, Beckmann CF. "Inter-individual differences in human brain structure and morphology link to variation in demographics and behavior". eLife. 2019;8:e44443. https://doi.org/10.7554/eLife.44443.
15. Karlsson Linnér R, et al. "Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences". Nature Genetics. 2019;51:245–57. https://doi.org/10.1038/s41588-018-0309-3.
16. Casey BJ, et al. "The Adolescent Brain Cognitive Development (ABCD) study: Imaging acquisition across 21 sites," Developmental Cognitive Neuroscience. 2018;32:43-54.
17. Loth E, et al. The EU-AIMS Longitudinal European Autism Project (LEAP): design and methodologies to identify and validate stratification biomarkers for autism spectrum disorders. Molecular Autism. 2017;8:24. https://doi.org/10.1186/s13229-017-0146-8.
18. Von Rhein D, et al. The NeuroIMAGE study: a prospective phenotypic, cognitive, genetic and MRI study in children with attention-deficit/hyperactivity disorder. Design and descriptives. Eur Child Adolesc Psychiatry. 2015;24(3):265–81. https://doi.org/10.1007/s00787-014-0573-4.
19. Murphy D, Spooren W. EU-AIMS: a boost to autism research. Nat Rev Drug Discovery. 2012;11(11):815–6. https://doi.org/10.1038/nrd3881.
20. Collins R. What makes UK Biobank special? The Lancet. 2012;379(9822):1173–4. https://doi.org/10.1016/S0140-6736(12)60404-8.
21. Van Essen DC, Smith SM, Barch DM, Behrens TEJ, Yacoub E, Ugurbil K. The WU-Minn human connectome project: an overview. Neuroimage. 2013;80:62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041.
22. Sudlow C, et al. "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age". PLoS Med. 2015;12(3):e1001779. https://doi.org/10.1371/journal.pmed.1001779.
23. Elam JS, Van Essen PD. "Human Connectome Project". In: Jaeger D, Jung R, editors. R. (eds) Encyclopedia of Computational Neuroscience. Neuroscience. New York: Springer; 2015. p. 1408–11.
24. Thompson PM, et al. ENIGMA and global neuroscience: A decade of large-scale studies of the brain in health and disease across more than 40 countries. Transl Psychiatry. 2020;10(1):100. https://doi.org/10.1038/s41398-020-0705-1.
25. Di Martino A, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol Psychiatry. 2014;19(6):659–67. https://doi.org/10.1038/mp.2013.78.
26. Little RJA. A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc. 1988;8(404):198-202. https://doi.org/10.1080/01621459.1988.10478722.
27. Little RJA, Rubin DB. Statistical analysis with missing data. Wiley; 2019. https://doi.org/10.1002/9781119482260.
28. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials - A practical guide with flowcharts. BMC Med Res Methodol. 2017;17(1):162. https://doi.org/10.1186/s12874-017-0442-1.
29. Golub GH, Van Loan CF. Matrix computations. 1996. MD, USA: Johns Hopkins Univ. Press. Balt; 1996.
30. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol. 2006;59(10):1087–91. https://doi.org/10.1016/j.jclinepi.2006.01.014.
31. Alexopoulos EC. "Introduction to multivariate regression analysis". Hippokratia. 2010;14(Suppl 1):23–8.
32. Hayati Rezvan P, Lee KJ, Simpson JA. "The rise of multiple imputation: A review of the reporting and implementation of the method in medical research Data collection, quality, and reporting". BMC Med Res Methodol. 2015;15(30). https://doi.org/10.1186/s12874-015-0022-1.
33. Kleinrock L. "Analysis of A time-shared processor". Naval Research Logistics Quarterly. 1964;11(1):59-73. https://doi.org/10.1002/nav.3800110105.
34. van Buuren S, Groothuis-Oudshoorn K. "mice: Multivariate imputation by chained equations in R". J Stat Softw. 2011;45(3):1–67. https://doi.org/10.18637/jss.v045.i03.
35. Zhu J, Raghunathan TE. "Convergence Properties of a Sequential Regression Multiple Imputation Algorithm". J Am Stat Assoc. 2015;110(511):1112-24. https://doi.org/10.1080/01621459.2014.948117.
36. Bartlett JW, Seaman SR, White IR, Carpenter JR. Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. Stat Methods Med Res. 2015;24(4):462–87. https://doi.org/10.1177/0962280214521348.
37. Stekhoven DJ, Bühlmann P. Missforest-Non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28(1):112–8. https://doi.org/10.1093/bioinformatics/btr597.
38. Lynn P, Rubin DB. "Multiple Imputation for Nonresponse in Surveys". J R Stat Soc: Ser D: The Statistician. 1988;37(4/5):475-6. https://doi.org/10.2307/2348774.
39. Charman T, et al. The EU-AIMS Longitudinal European Autism Project (LEAP): clinical characterisation. Molecular Autism. 2017;8:27. https://doi.org/10.1186/s13229-017-0145-9.
40. Lord C, et al. The Autism Diagnostic Observation Schedule-Generic: A standard measure of social and communication deficits associated with the spectrum of autism. J Autism Dev Disord. 2000;30(3):205-23. https://doi.org/10.1023/A:1005592401947.
41. Lord C, Rutter M, Le Couteur A. Autism diagnostic interview-revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. J Autism Dev Disord. 1994;24(5):659–85. https://doi.org/10.1007/BF02172145.
42. Sparrow SS. "Vineland Adaptive Behavior Scales". Encycl Clin Neuropsychol. 2011.
43. Constantino JN, Gruber CP. "Social responsiveness scale: SRS-2". Torrance: Western Psychological Services; 2012.
44. Bodfish JW, Symons FJ, Parker DE, Lewis MH. Varieties of repetitive behavior in autism: comparisons to mental retardation. J Autism Dev Disord. 2000;30(3):237-43. https://doi.org/10.1023/A:1005596502855.
45. Paterson H, Peck K, Perry KJ, Hickson M, Thomas J. The sensory profile: users manual. San Antonio; Psychological Corp. 1999.
46. Owens J, Maxim R, McGuinn M, Nobile C, Msall M, Alario A. "Television-viewing habits and sleep disturbance in school children". Pediatrics. 1999;104(3):e27. https://doi.org/10.1542/peds.104.3.e27.
47. Auyeung B, Wheelwright S, Allison C, Atkinson M, Samarawickrema N, Baron-Cohen S. The children's empathy quotient and systemizing quotient: sex differences in typical development and in autism spectrum conditions. Journal of Autism and Developmental Disorders. 2009;104(3):e27. https://doi.org/10.1007/s10803-009-0772-x.
48. Baron-Cohen S, Richler J, Bisarya D, Gurunathan N, Wheelwright S. The systemizing quotient: an investigation of adults with Asperger syndrome or high-functioning autism, and normal sex differences. Philos Trans R Soc B Biol Sci. 2003;358(1430):361–74. https://doi.org/10.1098/rstb.2002.1206.
49. Auyeung B, Allison C, Wheelwright S, Baron-Cohen S. Brief report: development of the adolescent empathy and systemizing quotients. J Autism Dev Disord. 2012;42(10):2225-35. https://doi.org/10.1007/s10803-012-1454-7.
50. Starfield B, et al. The adolescent child health and illness profile: a population-based measure of health. Med Care. 1995;33(5):553-66. https://doi.org/10.1097/00005650-199505000-00008.
51. Goodman R. The strengths and difficulties questionnaire: a research note. J Child Psychol Psychiatry Allied Discip. 1997;38(5):581-6. https://doi.org/10.1111/j.1469-7610.1997.tb01545.x.

52. Goodman R, Ford T, Richards H, Gatward R, Meltzer H. The development and well-being assessment: description and initial validation of an integrated assessement of child and adolescent psychopathology. J Child Psychol Psychiatry Allied Discip. 2000;41(5):645-55. https://doi.org/10.1017/S0021963099005909.
53. Beck AT, Steer RA. Beck Anxiety Inventory Manual. San Antonio: Harcourt Brace and Co; 1993.
54. Beck AT, Steer RA, Brown GK. Manual for the Beck depression inventory-II. San Antonio: The Psychological Corporation; 1996.
55. McCrimmon AW, Smith AD. Review of the Wechsler Abbreviated Scale of Intelligence, Second Edition (WASI-II). WechslerD (2011) Wechsler Abbreviated Scale of Intelligence, Second Edition (WASI-II). NCS Pearson. San Antonio: J Psychoeduc Assess; 2013;31(3):337-41.
56. Oldfield RC. The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia. 1971;9(1):97-113. https://doi.org/10.1016/0028-3932(71)90067-4.
57. Castelli F, Frith C, Happé F, Frith U. Autism, asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. Brain. 2002;125(8):1839–49. https://doi.org/10.1093/brain/awf189.
58. Senju A, Southgate V, White S, Frith U. "Mindblind eyes: an absence of spontaneous theory of mind in Asperger syndrome". Science. 2009;325(5942):883-5. https://doi.org/10.1126/science.1176170.
59. Sjöwall D, Roth L, Lindqvist S, Thorell LB. Multiple deficits in ADHD: executive dysfunction, delay aversion, reaction time variability, and emotional deficits. J Child Psychol Psychiatry Allied Discip. 2013. https://doi.org/10.1111/jcpp.12006.
60. Blasi G, et al. Differentiating allocation of resources and conflict detection within attentional control processing. Eur J Neurosci. 2007. https://doi.org/10.1111/j.1460-9568.2007.05283.x.
61. Baumeister S, et al. "Attenuated anticipation of social and monetary rewards in autism spectrum disorders". bioRxiv. 2020.07.06.186650, 2020. https://doi.org/10.1101/2020.07.06.186650.
62. Hariri AR, et al. "Serotonin transporter genetic variation and the response of the human amygdala," Science. 2002;297(5580):400-3. https://doi.org/10.1126/science.1071829.
63. White SJ, Coniston D, Rogers R, Frith U. Developing the Frith-Happé animations: a quick and objective test of Theory of Mind for adults with autism. Autism Res. 2011;4(2):149-54. https://doi.org/10.1002/aur.174.
64. Sambataro F, et al. Altered cerebral response during cognitive control: a potential indicator of genetic liability for schizophrenia. Neuropsychopharmacology. 2013;38(5):846-53. https://doi.org/10.1038/npp.2012.250.
65. Constantino JN, et al. Validation of a brief quantitative measure of autistic traits: comparison of the social responsiveness scale with the autism diagnostic interview-revised. J Autism Dev Disord. 2003;3(4):427-33. https://doi.org/10.1023/A:1025014929212.
66. De Bildt A, Mulder EJ, Hoekstra PJ, Van Lang NDJ, Minderaa RB, Hartman CA. Validity of the children's social behavior Qquestionnaire (CSBQ) in children with intellectual disability: comparing the CSBQ with ADI-R, ADOS, and clinical DSM-IV-TR classification. J Autism Dev Disord. 2009;39(10):1464-70. https://doi.org/10.1007/s10803-009-0764-x.
67. Hartman CA, Luteijn E, Serra M, Minderaa R. Refinement of the Children's Social Behavior Questionnaire (CSBQ): an instrument that describes the diverse problems seen in milder forms of PDD. J Autism Dev Disord. 2006;36(3):325-42. https://doi.org/10.1007/s10803-005-0072-z.
68. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med. 2011;30(4):377-99. https://doi.org/10.1002/sim.4067.
69. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. Am J Epidemiol. 2014;179(6):764-74. https://doi.org/10.1093/aje/kwt312.
70. LeCun Y, et al. Backpropagation applied to handwritten zip code recognition. Neural Comput. 1989;1:541-51. https://doi.org/10.1162/neco.1989.1.4.541.
71. Altman NS. "An introduction to kernel and nearest-neighbor nonparametric regression". The American Statistician. 1992;46(3):175-85. https://doi.org/10.1080/00031305.1992.10475879.
72. Mason D, McConachie H, Garland D, Petrou A, Rodgers J, Parr JR. Predictors of quality of life for autistic adults. Autism Res. 2018;11(8):1138–47. https://doi.org/10.1002/aur.1965.
73. Kuhlthau K, et al. Health-related quality of life for children with ASD: associations with behavioral characteristics. Res Autism Spectr Disord. 2013;40(6):721-9. https://doi.org/10.1016/j.rasd.2013.04.006.
74. Bishop CM. Pattern Recognition and Machine Learning (Information Science and Statistics), 1st ed. New York: Springer; 2007.
75. MacKay DJC. Information Theory, Inference and Learning Algorithms. Cambridge: Cambridge University Press; 2003.
76. Neal R. "Bayesian Learning for Neural Networks". Lecture notes in statistics. New York: Springer Verslag; 1996.
77. Gordon AD, Breiman L, Friedman JH, Olshen RA, Stone CJ. "Classification and Regression Trees,". Biometrics. 1984:874. https://doi.org/10.2307/2530946.
78. Geurts P, Ernst D, Wehenkel L. "Extremely randomized trees". Mach Learn. 2006;63:3–42. https://doi.org/10.1007/s10994-006-6226-1.
79. Yucel RM, Zaslavsky AM. Practical suggestions on rounding in multiple imputation. ASA Section on Survey Research Methods. 2001.
80. Pedregosa F, et al. "Scikit-learn: Machine learning in Python". J Mach Learn Res. 2011;12:2825-30.
81. Rubin DB. Inference and missing data. Biometrika. 1976. https://doi.org/10.1093/biomet/63.3.581.

Publisher's Note