Is there a role for statistics in artificial intelligence?

Sarah Friedrich, Gerd Antes, Sigrid Behr, Harald Binder, Werner Brannath, Florian Dumpert, Katja Ickstadt, Hans A. Kestler, Johannes Lederer, Heinz Leitgöb, Markus Pauly, Ansgar Steland, Adalbert Wilhelm, Tim Friede

REGULAR ARTICLE

Is there a role for statistics in artificial intelligence?

Sarah Friedrich, et al. *[full author details at the end of the article]*

Abstract

The research on and application of artificial intelligence (AI) has triggered a comprehensive scientific, economic, social and political discussion. Here we argue that statistics, as an interdisciplinary scientific field, plays a substantial role both for the theoretical and practical understanding of AI and for its future development. Statistics might even be considered a core element of AI. With its specialist knowledge of data evaluation, starting with the precise formulation of the research question and passing through a study design stage on to analysis and interpretation of the results, statistics is a natural partner for other disciplines in teaching, research and practice. This paper aims at highlighting the relevance of statistical methodology in the context of AI development. In particular, we discuss contributions of statistics to the field of artificial intelligence concerning methodological development, planning and design of studies, assessment of data quality and data collection, differentiation of causality and associations and assessment of uncertainty in results. Moreover, the paper also discusses the equally necessary and meaningful extensions of curricula in schools and universities to integrate statistical aspects into AI teaching.

1 Introduction

The research on and application of artificial intelligence (AI) has triggered a comprehensive scientific, economic, social and political discussion. Here we argue that statistics, as an interdisciplinary scientific field, plays a substantial role, both for the theoretical and practical understanding of AI and for its further development.

Contrary to the public perception, AI is not a new phenomenon. AI was already mentioned in 1956 at the Dartmouth Conference (Moor 2006; Solomonoff 1985), and the first data-driven algorithms such as Perceptron (Rosenblatt 1958), backpropagation (Kelley 1960) and the so-called 'Lernmatrix', an early neural system (Steinbuch 1961;

✉ Tim Friede
    tim.friede@med.uni-goettingen.de

Extended author information available on the last page of the article

Hilberg 1995), were developed in the 50s and 60s. The Lighthill Report in 1973 made a predominantly negative judgment on AI research in Great Britain and led to the fact that the financial support for AI research was almost completely stopped (the so-called *first AI winter*). The following phase of predominantly knowledge-based development ended in 1987 with the so-called *second AI winter*. A period of reduced public interest and funding in AI began. Nonetheless, in 1988, Judea Pearl published his book 'Probabilistic Reasoning in Intelligent Systems', for which he received the Turing Award in 2011 (Pearl 1988). From the beginning of the 1990s, AI has been developing again with major breakthroughs like Support Vector Machines (Cortes and Vapnik 1995), Random Forest (Breiman 2001), Bayesian Methods (Zhu et al. 2017), Boosting and Bagging (Freund and Schapire 1997; Breiman 1996), Deep Learning (Schmidhuber 2015) and Extreme Learning Machines (Huang et al. 2006).

Today, AI plays an increasingly important role in many areas of life. International organizations and national governments have currently positioned themselves or introduced new regulatory frameworks for AI. Examples are, among others, the AI strategy of the German government (Bundesregierung 2018), the statement of the Data Ethics Commission (Data Ethics Commission of the Federal Government, Federal Ministry of the Interior, Building and Community 2019) from 2019 and the report of the Nuffield Foundation in the UK (Nuffield Foundation 2019). Similarly, the European Commission recently published a white paper on AI (European Commission 2020b). Furthermore, regulatory authorities such as the US Food and Drug Administration (FDA) are now also dealing with AI topics and their evaluation. In 2018, for example, the electrocardiogram function of the Apple Watch was the first AI application to be approved by the FDA (MedTechIntelligence 2018).

There is no unique and comprehensive definition of *artificial intelligence*. Two concepts are commonly used distinguishing *weak* and *strong AI*. Searle (1980) defined them as follows: *'According to weak AI, the principal value of the computer in the study of the mind is that it gives us a very powerful tool. [...] But according to strong AI, the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really is a mind [...]'.* Thus, strong AI essentially describes a form of machine intelligence that is equal to human intelligence or even improves upon it, while weak AI (sometimes also referred to as narrow AI) is limited to tractable applications in specific domains. Following this definition, we will focus on weak AI in this paper in the sense that we consider self-learning systems, which are solving specific application problems based on methods from mathematics, statistics and computer science. Consequently, we will focus on the data-driven aspects of AI in this paper. In addition, there are many areas in AI that deal with the processing of and drawing inference from symbolic data (Bock and Diday 2000; Billard and Diday 2006). In contrast to standard data tables, symbolic data may consist of, e.g. lists, intervals, etc. Thus, special methods for data aggregation and analysis are necessary, which will not be discussed here.

As for AI, there is neither a single definition nor a uniform assignment of methods to the field of machine learning (ML) in literature and practice. Often, ML is considered a subset of AI approaches. Based on Simon's definition from 1983 (Simon 1983), learning describes changes of a system in such a way that a similar task can be performed more effectively or efficiently the next time it is performed. Bishop (2006)

describes machine learning as the *'automatic discovery of regularities in data through the use of computer algorithms [...]'*. Following these concepts, we use AI in a very general sense in this paper whereas ML is used to refer to more specific (statistical) algorithms.

Often the terms AI and ML are mentioned along with *Big Data* (Gudivada et al. 2015) or *Data Science*, sometimes even used interchangeably. However, neither are AI methods necessary to solve Big Data problems, nor are methods from AI only applicable to Big Data. Data Science, on the other hand, is usually considered as an intersection of computer science, statistics and the respective scientific discipline. Therefore, it is not bound to certain methods or certain data conditions.

This paper aims at contributing to the current discussion about AI by highlighting the relevance of statistical methodology in the context of AI development and application. Statistics can make important contributions to a more successful and secure use of AI systems, for example with regard to

1. Design (Sect. 3): bias reduction; validation; representativity; selection of variables
2. Assessment of data quality (Sect. 4): standards for the quality of diagnostic tests and audits; dealing with missing values
3. Differentiation between causality and associations (Sect. 5): consideration of covariate effects; answering causal questions; simulation of interventions
4. Assessment of certainty or uncertainty in results (Sect. 6): Increasing interpretability; mathematical validity proofs or theoretical properties in certain AI contexts; providing stochastic simulation designs; accurate analysis of the quality criteria of algorithms in the AI context

The remainder of the paper is organized as follows: First, we present an overview of AI applications and methods in Sect. 2. We continue by expanding on the points 1.–4. in Sects. 3, 4, 5 and 6. We conclude with Sect. 7. There, we also discuss the increased need for teaching and further education targeting the increase of AI-related literacy (particularly with respect to the underlying statistical concepts) at all educational levels.

2 Applications and methods of AI

Important categories of AI approaches are supervised learning, unsupervised learning and reinforcement learning (Sutton and Barto 2018). In supervised learning, AI systems learn from training data with known output such as true class labels or responses. Thus, the aim is to learn some function $g : X \rightarrow Y$ describing the relationship between an $n \times p$ matrix of given features $\mathbf{X} \subset X$ and the vector of labels $\mathbf{Y} = (y_1, \ldots, y_n)' \subset Y$. Here, n denotes the number of observations, p is the number of features and X and Y describe the input and output space, respectively. Examples include, among others, support-vector machines, linear and logistic regression or decision trees. In contrast, unsupervised learning extracts patterns from unlabeled data, i.e. without the y_is in the notation above. The most well-known examples include principal component analysis and clustering. Finally, reinforcement learning originates from robotics and describes the situation where an 'agent' (i.e. an autonomous entity with the ability to act and direct its activity towards achieving goals) learns through

trial-and-error search. Markov decision processes from probability theory play an important role here (Sutton and Barto 2018). The input data to an AI algorithm can be measured values such as stock market prices, audio signals, climate data or texts, but may also describe very complex relationships, such as chess games. In the following, we provide some specific examples of AI applications.

2.1 Applications of AI

AI has made remarkable progress in various fields of application. These include automated face recognition, automated speech recognition and translation (Barrachina et al. 2009), object tracking in film material, autonomous driving, and the field of strategy games such as chess or go, where computer programs now beat the best human players (Koch 2016; Silver et al. 2018).

Especially for tasks in speech recognition as well as text analysis and translation, Hidden Markov models from statistics are used and further developed with great success (Juang and Rabiner 1991; Kozielski et al. 2013) because they are capable of representing grammars. Nowadays, automatic language translation systems can even translate languages such as Chinese into languages of the European language family in real time and are used, for example, by the EU (European Commission 2020a). Another growing area for AI applications is medicine. Here, AI is used, e.g., to improve the early detection of diseases, for more accurate diagnoses, or to predict acute events (Burt et al. 2018; Chen et al. 2018), see also Friedrich et al. (2021) for a recent overview. Directions for future developments include personalized medicine aiming at tailoring treatments to patient subgroups (strata) or even individual patients (Hamburg and Collins 2010; Blasiak et al. 2020; Schork 2019). Furthermore, official statistics uses AI methods for classification as well as for recognition, estimation and/or imputation of relevant characteristic values of statistical units (Beck et al. 2018; Ramosaj and Pauly 2019b; Ramosaj et al. 2020; UNECE 2020; Thurow et al. 2021). In economics and econometrics, AI methods are also applied and further developed, for example, to draw conclusions about macroeconomic developments from large amounts of data on individual consumer behavior (McCracken and Ng 2016; Ng 2018).

Despite these positive developments that also dominate the public debate, some caution is advisable. There are a number of reports about the limits of AI, e.g., in the case of a fatal accident involving an autonomously driving vehicle (Wired.com 2019). Due to the potentially serious consequences of false positive or false negative decisions in AI applications, careful consideration of these systems is required (AInow 2020). This is especially true in applications such as video surveillance of public spaces. For instance, a pilot study conducted by the German Federal Police at the Südkreuz suburban railway station in Berlin has shown that automated facial recognition systems for identification of violent offenders currently have false acceptance rates of 0.67% (test phase 1) and 0.34% (test phase 2) on average (Bundespolizeipräsidium Potsdam 2018). This means that almost one in 150 (or one in 294) passers-by is falsely classified as a violent offender. In medicine, wrong decisions can also have drastic and negative effects, such as an unnecessary surgery and chemotherapy in the case of wrong cancer diagnoses. Corresponding test procedures for assessing such diagnostic tests
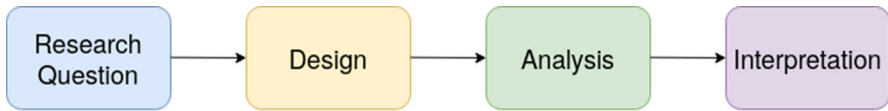
Fig. 1 Flow chart of study planning, design, analysis and interpretation

for medicine are currently being developed by regulators such as the US FDA (FDA 2019).

2.2 Methods for AI and the role of statistics

Even though many of the contributions to AI systems originate from computer science, statistics has played an important role throughout. Early examples occurred in the context of realizing the relationship between backpropagation and nonlinear least squares methods, see, e.g., Warner and Misra (1996). Important ML methods such as random forests (Breiman 2001) or support vector machines (Cortes and Vapnik 1995) were developed by statisticians. Others, like radial basis function networks (Chen et al. 1991), can also be considered and studied as nonlinear regression models in statistics. Recent developments such as extreme learning machines or broad learning systems (Chen and Liu 2018) have close links to multiple multivariate and ridge regression, i.e. to statistical methods. The theoretical validity of machine learning methods, e.g., through consistency statements and generalization bounds (Györfi et al. 2002; Vapnik 1998), also requires substantial knowledge of mathematical statistics and probability theory.

To capture the role and relevance of statistics, we consider the entire process of establishing an AI application. As illustrated in Fig. 1, various steps are necessary to examine a research question empirically. For more details on these steps see, e.g., Weihs and Ickstadt (2018). Starting with the precise formulation of the research question the process then runs through a study design stage (including sample size planning and bias control) to the mathematical formulation (e.g. as an optimization problem) and the numerical analysis. Finally, the results must be interpreted. AI often focuses on the step of data analysis while the other stages receive less attention or are even ignored. This may result in critical issues and possibly misleading interpretations, such as sampling bias or the application of inappropriate analysis tools requiring assumptions not met by the chosen design.

3 Statistical approaches for study design and validation

The design of a study and the data to be considered is the basis for the validity of the conclusions. Unfortunately, AI applications often use data that were collected for a different purpose (so-called secondary data, observational studies). The collection and compilation of secondary data is in general not based on a specific purpose or a research question. Instead, it is collected for other purposes such as accounting or storage purposes. A typical case is the scientific use of routine data. For example, the

AI models in a recent study about predicting medical events (such as hospitalization) are based on medical billing data (Lin et al. 2019). Another typical case concerns convenience samples, that is, samples that are not randomly drawn but instead depend on 'availability'. Well-known examples are online questionnaires, which only reach those people who visit the corresponding homepage and take the time to answer the questions. The concept of knowledge discovery in databases (Fayyad et al. 1996) very clearly reflects the assumption that data are regarded as a given basis from which information and knowledge can be extracted by AI procedures. This is contrary to the traditional empirical research process, in which empirically testable research questions are derived from theoretical questions by conceptualizing and operationalizing. Importantly, the resulting measurement variables are then collected for this specific purpose.

3.1 Validation

Statistics distinguishes between two types of validity (Shadish et al. 2002):

1. Internal validity is the ability to attribute a change in an outcome of a study to the investigated causes. In clinical research, e.g., this type of validity is ensured through randomization in controlled trials. Internal validity in the context of AI and ML can also refer to avoiding systematic bias (such as systematically underestimated risks).

2. External validity is the ability to transfer the observed effects and relationships to larger or different populations, environments, situations, etc. In the social sciences (e.g. in the context of opinion research), an attempt to achieve this type of validity is survey sampling, which comprises sampling methods that aim at representative samples in the sense of Gabler and Häder (2018), see also Kruskal and Mosteller (1979a, b, c, 1980).

These validation aspects are important, but different traditions exist for AI algorithms and statistics: While ML has a longstanding benchmarking tradition and often uses many datasets for evaluation, statistics tends to rely on theory and simulations augmented by one or two convincing data examples. Here, statistics makes use of probabilistic models in order to reflect a diversity of real life situations. In addition to mathematical validity proofs and theoretical investigations, detailed simulation studies are carried out to evaluate the methods' limits (by exceeding the assumptions made) and finite sample properties in situations where certain properties can only be proven asymptotically. This statistical perspective provides useful insights.

Concepts and guidelines for designing, structuring and reporting simulation studies have a longstanding tradition in medical statistics, see for example Burton et al. (2006), Friede et al. (2010), Benda et al. (2010), Morris et al. (2019).

A particular challenge for validation of AI systems are the ever faster development cycles which require continuous investigations. This can even be aggravated when turning to development processes of mobile apps or online learning systems such as recommender systems in online shopping portals. Here, the developments are dynamic, de facto never ending processes, which therefore require continuous validation.

Another important factor for the validity and reliability of a study is the sample size (Meng 2018). For high-dimensional models, an additional factor is 'sparsity': In many applications, the input data for the AI techniques are very high-dimensional, i.e. a large number of variables p (also called features) are observed with diverse ranges of possible values. In addition, non-linear relationships with complex interactions are often considered for prediction. It is well known that high-dimensional data are difficult to handle if sample sizes are small, i.e. if $p \gg n$. However, even with sample sizes in the order of millions, the problem of the *curse of dimensionality* arises (Bellman 1957), because data is thin and sparse in a high-dimensional space, i.e. only few variables are related to the outcome. Therefore, learning the structure of the high-dimensional space from these thin data typically requires an enormous amount of training data. Through statistical models and corresponding mathematical approximations or numerical simulations, statisticians can assess the potentials and limits of an AI application for a given number of cases or estimate the necessary number of cases in the planning stage of a study. This is not routine work; instead, it requires advanced statistical training, competence, and experience.

Thus, statistics can help in collecting and processing data for subsequent use in AI pipelines. Basic statistical techniques that are relevant for this aspect include, for example, the modeling of the data generating process, restrictions on data sets (Rubin 2008), and factorial design of experiments, which is a controlled variation of factors highlighting their respective influence. In addition, the various phases in the development of a diagnostic test are well known in statistics (Pepe 2003), with (external) validation on independent data playing a crucial role. In many AI applications, however, the final evaluation phase on external data is never reached, since the initial algorithms have been replaced in the meantime. Also, statistical measures of quality such as sensitivity, specificity, ROC curves and calibration are used in the evaluation of AI methods. And finally, statistics can help in the assessment of uncertainty (Sect. 6).

3.2 Representativity

The naive expectation that sufficiently large data automatically leads to representativity is wrong (Meng 2018; Meng and Xie 2014). A prominent example is Google Flu (Lazer et al. 2014), where flu outbreaks were predicted on the basis of search queries: it turned out that the actual prevalence of the flu was overestimated considerably. Another example is Microsoft's chatbot Tay (Davis 2016; Wolf et al. 2017), which was designed to mimic the speech pattern of a 19-year-old American girl and to learn from interactions with human users on Twitter: after only a short time, the bot posted offensive and insulting tweets, forcing Microsoft to shut down the service just 16 hours after it started. And yet another example is the recently published Apple Heart Study (Perez et al. 2019), which examined the ability of Apple Watch to detect atrial fibrillation: there were more than 400,000 participants, but the average age was 41 years, which is particularly problematic in view of atrial fibrillation occurring almost exclusively in people over 65 years of age.
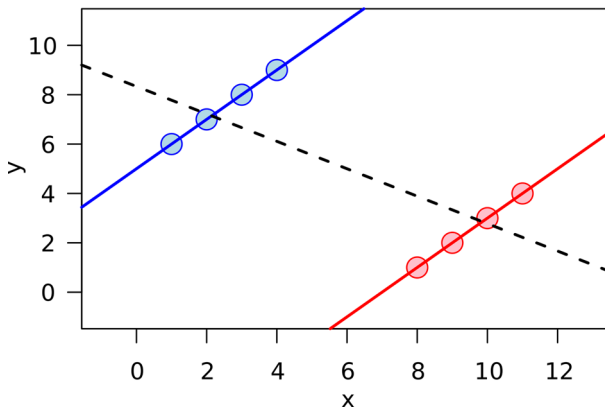
Fig. 2 Simpson's paradox for continuous data: a positive trend is visible for both groups individually (red and blue), but a negative trend (dashed line) appears when the data are pooled across groups (Wikipedia 2020) (color figure online)

3.3 Bias

If careful data collection is not accounted for, spurious correlations and bias can falsify the conclusions. Many forms of bias exist, such as selection, attribution, performance, and detection bias. While bias in the context of statistics usually refers to the deviation between the estimated and the true value of a parameter, there are also other concepts such as cognitive biases or, as Ntoutsi et al. (2020) put it *'the inclination or prejudice of a decision made by an AI system which is for or against one person or group, especially in a way considered to be unfair'*. A classic example of such falsification is Simpson's paradox (Simpson 1951), which describes a reversal of group-specific trends when subgroups are disregarded, see Fig. 2. Further examples are biases inflicted by how the data are collected, such as length time bias (Porta 2016) or prejudices introduced by AI, see Ntoutsi et al. (2020) for a recent overview of this topic.

Statistics provides methods and principles for minimizing bias. Examples include the assessment of the risk of bias in medicine (Higgins et al. 2011), stratification, marginal analyses, consideration of interactions, meta-analyses, and techniques specifically designed for data collection such as (partial) randomization, (partial) blinding, and methods of so-called optimal designs (Karlin and Studden 1966). Statistics also provides designs that allow for the verification of internal and external validity (Bartels et al. 2018; Braver and Smith 1996; Roe and Just 2009).

3.4 Model stability and reproducibility

Whether there is interest in a model for prediction or in a descriptive model, model stability, i.e. the robustness of the model towards small changes in the input values, plays an important role. Variable selection methods are used to derive descriptive models and model complexity has an important influence on the choice of the methods. In a recent review of variable selection procedures, Heinze et al. (2018) emphasize the

Fig. 3 Data relevancy and quality are equivalent components of a fit-for-purpose real-world data set. Figure according to Duke-Margolis (2018)

important role of stability investigations. This issue is also mentioned in Sauerbrei et al. (2020) as one of the main target parameters for the comparison of variable selection strategies. Statistical concepts have been introduced by Meinshausen and Bühlmann (2010) as well as in the context of random forests (Breiman 2001).

Conscientiously sticking to the principles mentioned above and adhering to a previously defined study design also counteracts the so-called replication crisis (Pashler and Wagenmakers 2012). In this methodological crisis, which has been ongoing since the beginning of the 2010s, it has become clear that many studies, especially in medicine and the social sciences, are difficult or impossible to reproduce. Since reproducibility of experimental results is an essential part of scientific methodology (Staddon 2017), an inability to replicate the studies of others can have grave consequences for many fields of science. The replication crisis has been particularly widely discussed in psychology and medicine, where a number of efforts have been made to re-investigate previous findings in order to determine their reliability (Begley and Ellis 2012; Makel et al. 2012). A related issue is transparency. While this is an important concept in any empirical analysis, it has especially become an issue discussed in the context of AI applications, see for example Flake and Fried (2020), Haibe-Kains et al. (2020), Simons et al. (2017).

4 Statistics for the assessment of data quality

'Data is the new oil of the global economy.' According to, e.g., the New York Times (New York Times 2018) or the Economist (The Economist 2017), this credo echoes incessantly through start-up conferences and founder forums. This metaphor is not only popular but false. First of all, data in this context corresponds to crude oil, which needs further refining before it can be used. In addition, the resource crude oil is limited. 'For a start, while oil is a finite resource, data is effectively infinitely durable and reusable' [Bernard Marr in Forbes (2018)]. All the more important is a responsible approach to data preprocessing (Fig. 3).

Ensuring data quality is of great importance in all analyses, according to the popular slogan 'Garbage in, garbage out.' As already mentioned in the previous section, we mainly use secondary data in the context of AI. In AI, the process of operationalization is often replaced by the ETL process: 'Extract, Transform, Load' (Theodorou et al. 2017). Relevant measurements are to be extracted from the data lake(s), then transformed and finally loaded into the (automated) analysis procedures. Many AI procedures are thereby expected to be able to distill relevant influencing variables from high-dimensional data.

The success of this procedure fundamentally depends on the quality of the data. In line with Karr et al. (2006), data quality is defined here as the ability of data to be used quickly, economically and effectively for decision-making and evaluation (Karr et al. 2006). In this sense, data quality is a multi-dimensional concept that goes far beyond measurement accuracy and includes aspects such as relevance, completeness, availability, timeliness, meta-information, documentation and, above all, context-dependent expertise (Duke-Margolis 2018, 2019). In official statistics, relevance, accuracy and reliability, timeliness and punctuality, coherence and comparability, accessibility and clarity are defined as dimensions of data quality (European Statistical System 2019).

Increasing automation of data collection, e.g., through sensor technology, may increase measurement accuracy in a cost-effective and simple way. Whether this will achieve the expected improvement in data quality remains to be checked in each individual application. Missing values are a common problem of data analyses. In statistics, a variety of methods have been developed to deal with these, including imputation procedures, or methods of data enhancement (Rubin 1976; Seaman and White 2013; Van Buuren 2018). The AI approach of ubiquitous data collection allows the existence of redundant data, which can be used in a preprocessing step with appropriate context knowledge to complete incomplete data sets. However, this requires a corresponding integration of context knowledge into the data extraction process.

The data-hungry decision-making processes of AI and statistics are subject to a high risk with regard to relevance and timeliness, since they are implicitly based on the assumption that the patterns hidden in the data should perpetuate themselves in the future. In many applications, this leads to an undesirable entrenchment of existing stereotypes and resulting disadvantages, e.g., in the automatic granting of credit or the automatic selection of applicants. A specific example is given by the gender bias in Amazon's AI recruiting tool (Dastin 2018).

In the triad 'experiment - observational study - convenience sample (data lake)', the field of AI, with regard to its data basis, is moving further and further away from the classical ideal of controlled experimental data collection to an exploration of given data based on pure associations. However, only controlled experimental designs guarantee an investigation of causal questions. This topic will be discussed in more detail in Sect. 5. Causality is crucial if the aim of the analysis is to explain relationships such as the function $g : X \rightarrow Y$ linking the feature vector $\{x_1, \ldots, x_n\} \subset X$ to the outcome $\{y_1, \ldots, y_n\} \subset Y$. There are, however, other situations where one might not be primarily interested in causal conclusions. Good prediction, for example, can also be obtained by using variables that are not themselves causally related to the outcome but strongly correlated with some causal predictor instead.

Exploratory data analysis (Tukey 1962) provides a broad spectrum of tools to visualize the empirical distributions of the data and to derive corresponding key figures. This can be used in preprocessing to detect anomalies or to define ranges of typical values in order to correct input or measurement errors and to determine standard values. In combination with standardization in data storage, data errors in the measurement process can be detected and corrected at an early stage. This way, statistics helps to assess data quality with regard to systematic, standardized and complete recording. Survey methodology primarily focuses on data quality. The insights gained in statisti-

cal survey research to ensure data quality with regard to internal and external validity provide a profound foundation for corresponding developments in the context of AI. Furthermore, various procedures for imputing missing data are known in statistics, which can be used to complete the data depending on the existing context and expertise (Rubin 1976; Seaman and White 2013; Van Buuren 2018). Statisticians have dealt intensively with the treatment of missing values under different development processes [non-response, missing not at random, missing at random, missing completely at random (Rubin 1976; Molenberghs et al. 2014)], selection bias and measurement error (Keogh et al. 2020; Shaw et al. 2020).

Another point worth mentioning is parameter tuning, i.e. the determination of so-called hyperparameters, which control the learning behavior of ML algorithms: comprehensive parameter tuning of methods in the AI context often requires very large amounts of data. For smaller data volumes it is almost impossible to use such procedures. However, certain model-based (statistical) methods can still be used in this case (Richter et al. 2019).

5 Distinguishing between causality and association

Only a few decades ago, the greatest challenge of AI research was to program machines to associate a potential cause to a set of observable feature values, e.g. through Bayesian networks (Pearl 1988). The rapid development of AI in recent years (both in terms of the theory and methodology of statistical learning processes and the computing power of computers) has led to a multitude of algorithms and methods that have now mastered this task. One example are deep learning methods, which are used in robotics (Levine et al. 2018) and autonomous driving (Teichmann et al. 2018), as well as in computer-aided detection and diagnostic systems [e.g., for breast cancer diagnosis (Burt et al. 2018)], drug discovery in pharmaceutical research (Chen et al. 2018) and agriculture (Kamilaris and Prenafeta-Boldú 2018). With their often high predictive power, AI methods can uncover structures and relationships in large volumes of data based on associations. Due to the excellent performance of AI methods in large data sets, they are also frequently used in medicine to analyze register and observational data that have not been collected within the strict framework of a randomized study design (Sect. 3). However, the discovery of correlations and associations (especially in this context) is not equivalent to establishing causal claims.

An important step in the further development of AI is therefore to replace associational argumentation with causal argumentation. Pearl (2010) describes the difference as follows: *'An associational concept is any relationship that can be defined in terms of a joint distribution of observed variables, and a causal concept is any relationship that cannot be defined from the distribution alone.'*

Even the formal definition of a causal effect is not trivial. The fields of statistics and clinical epidemiology, for example, use the Bradford Hill criteria (Hill 1965) and the counterfactual framework introduced by Rubin (1974). The central problem in observational data are covariate effects, which, in contrast to the randomized controlled trial, are not excluded by design and whose (non-)consideration leads to distorted estimates of causal effects. In this context, a distinction must be made between *confounders*,
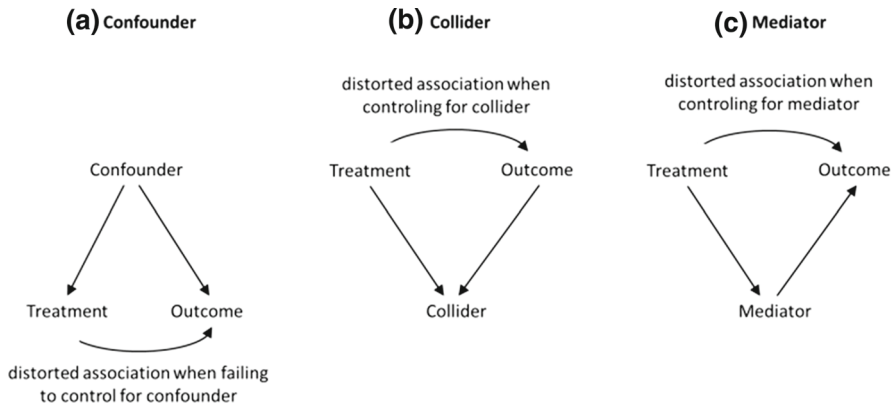
Fig. 4 Covariate effects in observational data, according to Catalogue of bias collaboration (2019)

colliders, and *mediators* (Pearl 2009). Confounders are unobserved or unconsidered variables that influence both the exposure and the outcome, see Fig. 4a. This can distort the effects of exposure if naively correlated. Fisher identified this problem in his book 'The Design of Experiments' published in 1935. A formal definition was developed in the field of epidemiology in the 1980s (Greenland and Robins 1986). Later, graphical criteria such as the Back-Door Criterion (Greenland et al. 1999; Pearl 1993) were developed to define the term *confounding*.

In statistics, the problem of confounding is taken into account either in the design (e.g., randomized study, stratification, etc.) or evaluation [propensity score methods (Cochran and Rubin 1973), marginal structural models (Robins et al. 2000), graphical models (Didelez 2007)]. In this context, it is interesting to note that randomized studies (which have a long tradition in the medical field) have recently been increasingly used in econometric studies (Athey and Imbens 2017; Duflo et al. 2007; Kohavi et al. 2020). In the case of observational data, econometrics has made many methodological contributions to the identification of treatment effects, e.g., via the potential outcome approach (Rosenbaum 2017, 2002, 2010; Rubin 1974, 2006) as well as the work on policy evaluation (Heckman 2001).

In contrast to confounders, colliders and mediators lead to distorted estimates of causal effects precisely when they are taken into account during estimation. Whereas colliders represent common consequences of treatment and outcome (Fig. 4b), mediators are variables that represent part of the causal mechanism by which the treatment affects the outcome (Fig. 4c). Especially in the case of longitudinal data, it is therefore necessary to differentiate in a theoretically informed manner which relationships the covariates in the observed data have with the treatment and outcome variables, thus avoiding bias in the causal effect estimates by (not) having taken them into account.

By integrating appropriate statistical theories and methods into AI, it will be possible to answer causal questions and simulate interventions. In medicine, e.g., questions such as 'What would be the effect of a general smoking ban on the German health care system' could then be investigated and reliable statements could be made, even without randomized studies which would not be possible here. Pearl's idea goes beyond

the use of ML methods in causal analyses (which are used, for example, in connection with targeted learning (Van der Laan and Rose 2011) or causal random forest (Athey and Imbens 2015)). His vision is rather to integrate the causal framework (Pearl 2010) described by him with ML algorithms to enable the machines to draw causal conclusions and simulate interventions.

The integration of statistical methods to detect causality in AI also contributes to increasing its transparency and thus the acceptance of AI methods, since a reference to probabilities or statistical correlations in the context of an explanation is not as effective as a reference to causes and causal effects (Miller 2019).

6 Statistical approaches for evaluating uncertainty and interpretability

Uncertainty quantification is often neglected in AI applications. One reason may be the above discussed misconception that 'Big Data' automatically leads to exact results, making uncertainty quantification redundant. Another key reason is the complexity of the methods which hampers the construction of statistically valid uncertainty assessments. However, most statisticians would agree that any comprehensive data analysis should contain methods to quantify the uncertainty of estimates and predictions. Its importance is also stressed by the American statistician David B. Dunson who writes that: *'it is crucial to not over-state the results and appropriately characterize the (often immense) uncertainty to avoid flooding the scientific literature with false findings.'* (Dunson 2018).

In fact, in order to achieve the main goal of highly accurate predictions, assumptions about underlying distributions and functional relationships are deliberately dropped in AI applications. On the one hand, this allows for a greater flexibility of the procedures. On the other hand, however, this also complicates an accurate quantification of uncertainty, e.g., to specify valid prediction and confidence regions for target variables and parameters of interest. As Bühlmann and colleagues put it: *'The statistical theory serves as guard against cheating with data: you cannot beat the uncertainty principle.'* (Bühlmann and van de Geer 2018). In recent years, proposals for uncertainty quantification in AI methods have already been developed by invoking Bayesian approximations, bootstrapping, jackknifing and other cross-validation techniques, Gaussian processes, Monte Carlo dropout etc., see e. g., Gal and Ghahramani (2016), Garnelo et al. (2018), Osband et al. (2016), Srivastava et al. (2014), Wager et al. (2014). However, their theoretical validity (e.g., that a prediction interval actually covers future values 95% of the time) has either not been demonstrated yet or has only been proven under very restrictive or at least partially unrealistic assumptions.

In contrast, algorithmic methods could be embedded in statistical models. While potentially less flexible, they permit a better quantification of the underlying uncertainty by specifying valid prediction and confidence intervals or allow for a better interpretation of the results. We give two examples: In time-to-event analyses mathematically valid simultaneous confidence bands for cumulative incidence functions can be constructed by combinations of nonparametric estimators of Kaplan-Meier or Aalen-Johansen-type and algorithmic resampling (Bluhmki et al. 2018; Dobler et al.

2017). Similarly, in the context of time series prediction, hybrid combinations of artificial neural networks with ARIMA models or within hierarchical structures allow for better explainability (Aburto and Weber 2007; Wickramasuriya et al. 2019).

Moreover, the estimated parameters of many AI approaches (such as deep learning) are difficult to interpret. Pioneering work from computer science on this topic is, for example, Valiant (1984, 2013), for which Leslie Valiant was awarded the Turing Award in 2010. Further research is nevertheless needed to improve interpretability. This also includes uncertainty quantification of patterns identified by an AI method, which heavily rely on statistical techniques. A tempting approach to achieve more interpretable AI methods is the use of auxiliary models. These are comparatively simple statistical models which, after adaptation of a deep learning approach, describe the most important patterns represented by the AI method and potentially can also be used to quantify uncertainty (Molnar 2019; Peltola 2018; Ribeiro et al. 2016a, b). In fact, as in computational and statistical learning theory (Györfi et al. 2002; Kearns and Vazirani 1994; Vapnik 1998), statistical methods and AI learning approaches can (and should) complement each other. Another important aspect is the model complexity which can, e.g., be captured by entropies (such as VC dimensions) or compression barriers (Langford 2005). These concepts as well as different forms of regularization (Tibshirani 1996; Wager et al. 2013; Zaremba et al. 2014), i.e. the restriction of the parameter space, allow to recognize or even to correct an overfitting of a learning procedure. Here, the application of complexity reducing concepts can be seen as a direct implementation of the Lex Parsimoniae principle and often increases the interpretability of resulting models (Ross et al. 2017; Tibshirani 1997). In fact, regularization and complexity reducing concepts are an integral part of many AI methods. However, they are also basic principles of modern statistics, which were already proposed before their introduction to AI. Examples are given in connection with empirical Bayesian or shrinkage methods (Röver and Friede 2020). In addition to that, AI and statistics have numerous concepts in common which give rise to an exchange of methods in these fields.

Furthermore, uncertainty aspects also apply to quality criteria (e.g., accuracy, sensitivity and specificity) of AI algorithms. The corresponding estimators are also random but their uncertainty is usually not quantified at all.

Statistics can help to increase the validity and interpretability of AI methods by providing contributions to the quantification of uncertainty. To achieve this, we can assume specific probabilistic and statistical models or dependency structures which allow comprehensive mathematical investigations (Athey et al. 2019; Bartlett et al. 2004; Devroye et al. 2013; Györfi et al. 2002; Scornet et al. 2015; Wager and Athey 2018; Ramosaj and Pauly 2019a), e.g., by investigating robustness properties, proving asymptotic consistency or (finite) error bounds. On the other hand this also includes the elaboration of (stochastic) simulation designs (Morris et al. 2019) and the specification of easy to interpret auxiliary statistical models. Finally, it allows for a detailed analysis of quality criteria of AI algorithms.

7 Conclusion and discussion

AI has been a growing research area for years, and its development will probably continue in the coming decades. In addition to ethical and legal problems, there are still many open questions regarding the collection and processing of data. Statistical methods must be considered as integral part of AI systems, from the formulation of the research questions, the development of the research design, through the analysis up to the interpretation of the results. Particularly in the field of methodological development, statistics can, e.g., serve as multiplier and strengthen the scientific exchange by establishing broad and strongly interconnected networks between users and developers.

In the context of clinical trials, statistics also provides guidelines for important aspects of trial design, data analysis and reporting. Many of these guidelines are currently being extended for AI applications, e.g. the TRIPOD statement (Collins et al. 2015; Collins and Moons 2019) or the CONSORT and SPIRIT guidelines (Liu et al. 2020; Rivera et al. 2020). Moreover, initiatives such as STRATOS (STRengthening Analytical Thinking for Observational Studies, https://stratos-initiative.org/) aim to provide guidance for applied statisticians and other data analysts with varying levels of statistical education.

As a core element of AI, statistics is the natural partner for other disciplines in teaching, research and practice. Therefore, it is advisable to incorporate statistical aspects into AI teaching and to bridge the gap between the two disciplines. This begins with school education, where statistics and computer science should be integral elements of the curricula, and continues with higher education as well as professional development and training. By developing professional networks, participating methodologists can be brought together with users/experts to establish or maintain a continuous exchange between the disciplines. In addition to AI methods, these events should also cover the topics of data curation, management of data quality and data integration.

Statistics is a broad cross-scientific discipline. Statisticians provide knowledge and experience of all aspects of data evaluation: starting with the research question through design and analysis to the interpretation. In particular, the following contributions of statistics to the field of artificial intelligence can be summarized:

1. Methodological development: The development of AI systems and their theoretical underpinning has benefited greatly from research in computer science and statistics, and many procedures have been developed by statisticians. Recent advances such as extreme learning machines show that statistics also provides important contributions to the design of AI systems, for example, by improved learning algorithms based on penalized or robust estimation methods.
2. Planning and design: Statistics can help to optimize data collection or preparation (sample size, sampling design, weighting, restriction of the data set, design of experiments, etc.) for subsequent evaluation with AI methods. Furthermore, the quality measures of statistics and their associated inference methods can help in the evaluation of AI models.
3. Assessment of data quality and data collection: Exploratory data analysis provides a wide range of tools to visualize the empirical distribution of the data and to derive

appropriate metrics, which can be used to detect anomalies or to define ranges of typical values, to correct input errors, to determine norm values and to impute missing values. In combination with standardization in data storage, errors in the measurement process can be detected and corrected at an early stage. With the help of model-based statistical methods, comprehensive parameter tuning is also possible, even for small data sets.

4. Differentiation of causality and associations: In statistics, methods for dealing with covariate effects are known. Here, it is important to differentiate theoretically informed between the different relationships covariates can have to treatment and outcome in order to avoid bias in the estimation of causal effects. Pearl's causal framework enables the analysis of causal effects and the simulation of interventions. The integration of causal methods into AI can also contribute to the transparency and acceptance of AI methods.

5. Assessment of certainty or uncertainty in results: Statistics can help to enable or improve the quantification of uncertainty in and the interpretability of AI methods. By adopting specific statistical models, mathematical proofs of validity can also be provided. In addition, limitations of the methods can be explored through (stochastic) simulation designs.

6. Conscientious implementation of points 2 to 5, including a previously defined evaluation plan, also counteracts the replication crisis (Pashler and Wagenmakers 2012) in many scientific disciplines. This aspect does not only hold for AI applications, but generally concerns all empirical studies.

7. Education, advanced vocational training and public relations: With its specialized knowledge, statistics is the natural partner for other disciplines in teaching and training. Especially in the further development of methods of artificial intelligence, statistics can strengthen scientific exchange.

With respect to some points raised in this paper, a few comments are in place. First, as mentioned in the introduction there is no unique definition of AI or ML according to the literature and distinguishing between the two is not easy. A broader consensus in the scientific community is necessary to facilitate common discussions. Second, as an anonymous referee commented, it might be helpful to distinguish between different frameworks concerning data and problems. The proposal is to distinguish between (a) problems and data with random or partly random aspects and (b) problems with a deterministic background such as graph theoretical structures or optimum configurations. While the first is a natural field of application for statistics, the second may also benefit from statistical approaches, e.g. concerning robustness or sensitivity. A related issue concerns the fact that the evaluation of AI methods must be seen in the context of the corresponding application. In life sciences and medicine we often assume the existence of some underlying 'ground truth' which needs to be estimated. Thus, modeling concepts such as bias or accuracy can be used for evaluation. In other areas such as economy or marketing, the idea rather is to derive a somewhat 'useful' or 'effective' strategy from the data. In such a situation, statistics can still be used for evaluation, for example by making predictions and comparing their accuracy with the observed data.

Another important aspect concerns the combination of data and results obtained from different studies. In evidence based medicine, systematic reviews and meta-analyses play a key role in combining results from multiple studies to give a quantitative summary of the literature. In contrast, meta-analysis methods to combine results from AI applications have not been developed yet. Initiatives to enable the sharing of data and models in AI include federated learning and software frameworks such as DataSHIELD (DataSHIELD 2018; Gaye et al. 2014), which enables remote and nondisclosive analysis of sensitive data, see also Bonofiglio et al. (2020). Thus, both fields could profit from an exchange of methods in this context.

The objective of statistics related to AI must be to facilitate or enable the interpretation of data. As Pearl puts it: *'Data alone are hardly a science, regardless how big they get and how skillfully they are manipulated'* (Pearl 2018). What is important is the knowledge gained that will enable future interventions.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Aburto L, Weber R (2007) Improved supply chain management based on hybrid demand forecasts. Appl Soft Comput 7(1):136–144

AInow (2020) https://ainowinstitute.org/, accessed 02.02.2020

Athey S, Imbens GW (2015) Machine learning for estimating heterogeneous causal effects. Stanford University, Graduate School of Business, Tech. rep

Athey S, Imbens GW (2017) The econometrics of randomized experiments. Handbook of Economic Field Experiments, vol 1. Elsevier, Amsterdam, pp 73–140

Athey S, Tibshirani J, Wager S (2019) Generalized random forests. Ann Stat 47(2):1148–1178

Barrachina S, Bender O, Casacuberta F, Civera J, Cubel E, Khadivi S, Lagarda A, Ney H, Tomás J, Vidal E, Vilar JM (2009) Statistical approaches to computer-assisted translation. Comput Linguistics 35(1):3–28. https://doi.org/10.1162/coli.2008.07-055-r2-06-29

Bartels DM, Hastie R, Urminsky O (2018) Connecting laboratory and field research in judgment and decision making: causality and the breadth of external validity. J Appl Res Memory Cogn 7(1):11–15. https://doi.org/10.1016/j.jarmac.2018.01.001

Bartlett PL, Bickel PJ, Bühlmann P, Freund Y, Friedman J, Hastie T, Jiang W, Jordan MJ, Koltchinskii V, Lugosi G et al (2004) Discussions of boosting papers, and rejoinders. Ann Stat 32(1):85–134

Beck M, Dumpert F, Feuerhake J (2018) Machine Learning in Official Statistics. arXiv preprint arXiv:1812.10422

Begley CG, Ellis LM (2012) Raise standards for preclinical cancer research. Nature 483(7391):531–533

Bellman R (1957) Dynamic programming. Princeton University Press, Princeton, New Jersey

Benda N, Branson M, Maurer W, Friede T (2010) Aspects of modernizing drug development using clinical scenario planning and evaluation. Drug Inf J DIJ/Drug Inf Assoc 44(3):299–315

Billard L, Diday E (2006) Symbolic data analysis: conceptual statistics and data mining. Wiley, Chichester, West Sussex

Bishop CM (2006) Pattern recognition and machine learning. Springer, New York

Blasiak A, Khong J, Kee T (2020) CURATE.AI: optimizing personalized medicine with artificial intelligence. SLAS TECHNOLOGY: Trans Life Sci Innov 25(2):95–105

Bluhmki T, Schmoor C, Dobler D, Pauly M, Finke J, Schumacher M, Beyersmann J (2018) A wild bootstrap approach for the Aalen–Johansen estimator. Biometrics 74(3):977–985

Bock HH, Diday E (2000) Analysis of symbolic data. Springer, Heidelberg

Bonofiglio F, Schumacher M, Binder H (2020) Recovery of original individual person data (ipd) inferences from empirical ipd summaries only: applications to distributed computing under disclosure constraints. Stat Med 39(8):1183–1198

Braver SL, Smith MC (1996) Maximizing both external and internal validity in longitudinal true experiments with voluntary treatments: The "combined modified" design. Eval Prog Planning 19(4):287–300. https://doi.org/10.1016/s0149-7189(96)00029-8

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140. https://doi.org/10.1007/bf00058655

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Bühlmann P, van de Geer S (2018) Statistics for big data: A perspective. Stat Prob Lett 136:37–41

Bundespolizeipräsidium Potsdam (2018) Abschlussbericht Teilprojekt 1 "Biometrische Gesichtserkennung". https://www.bundespolizei.de/Web/DE/04Aktuelles/01Meldungen/2018/10/181011_abschlussbericht_gesichtserkennung_down.pdf?__blob=publicationFile=1, accessed 07.05.2020

Bundesregierung (2018) Artificial intelligence strategy. https://www.ki-strategie-deutschland.de/home.html?file=files/downloads/Nationale_KI-Strategie_engl.pdf, accessed 07.05.2020

Burt JR, Torosdagli N, Khosravan N, RaviPrakash H, Mortazi A, Tissavirasingham F, Hussein S, Bagci U (2018) Deep learning beyond cats and dogs: recent advances in diagnosing breast cancer with deep neural networks. British J Radiol 91(1089):20170545

Burton A, Altman DG, Royston P, Holder RL (2006) The design of simulation studies in medical statistics. Stat Med 25(24):4279–4292

Catalogue of bias collaboration, Lee H, Aronson JK, Nunan D (2019) Catalogue of bias: Collider bias. https://catalogofbias.org/biases/collider-bias, accessed 12.02.2020

Chen CLP, Liu Z (2018) Broad learning system: an effective and efficient incremental learning system without the need for deep architecture. IEEE Trans Neural Netw Learn Syst 29(1):10–24. https://doi.org/10.1109/tnnls.2017.2716952

Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T (2018) The rise of deep learning in drug discovery. Drug Discovery Today 23(6):1241–1250

Chen S, Cowan CFN, Grant PM (1991) Orthogonal least squares learning algorithm for radial basis function networks. IEEE Trans Neural Netw 2(2):302–309. https://doi.org/10.1109/72.80341

Cochran WG, Rubin DB (1973) Controlling bias in observational studies: A review. Sankhyā: The Ind J Stat Ser A 35(4):417–446

Collins GS, Moons KG (2019) Reporting of artificial intelligence prediction models. The Lancet 393(10181):1577–1579

Collins GS, Reitsma JB, Altman DG, Moons KG (2015) Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) the TRIPOD statement. Circulation 131(2):211–219

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297. https://doi.org/10.1007/bf00994018

Dastin J (2018) Amazon scraps secret AI recruiting tool that showed bias against women. Reuters (2018). https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G, accessed 27.11.2019

Data Ethics Commission of the Federal Government, Federal Ministry of the Interior, Building and Community (2019) Opinion of the data ethics commission. https://www.bmi.bund.de/SharedDocs/downloads/EN/themen/it-digital-policy/datenethikkommission-abschlussgutachten-lang.pdf?__blob=publicationFile&v=4, accessed 07.05.2020

DataSHIELD (2018) https://www.datashield.ac.uk

Davis E (2016) AI amusements: the tragic tale of Tay the chatbot. AI Matters 2(4):20–24

Devroye L, Györfi L, Lugosi G (2013) A probabilistic theory of pattern recognition, vol 31. Springer, New York

Didelez V (2007) Graphical models for composable finite Markov processes. Scand J Stat 34(1):169–185

Dobler D, Beyersmann J, Pauly M (2017) Non-strange weird resampling for complex survival data. Biometrika 104(3):699–711

Duflo E, Glennerster R, Kremer M (2007) Using randomization in development economics research: A toolkit. Handbook of development economics 4:3895–3962

Duke-Margolis (2018) https://healthpolicy.duke.edu/sites/default/files/2020-03/characterizing_rwd.pdf. Accessed 13 May 2020

Duke-Margolis (2019) https://healthpolicy.duke.edu/sites/default/files/2019-11/rwd_reliability.pdf. Accessed 13 May 2020

Dunson DB (2018) Statistics in the big data era: Failures of the machine. Stat Prob Lett 136:4–9

European Commission (2020a) https://ec.europa.eu/info/resources-partners/machine-translation-public-administrations-etranslation_en#translateonline, accessed 13.05.2020

European Commission (2020b) On Artificial Intelligence - A European approach to excellence and trust. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf, accessed 29.07.2020

European Statistical System (2019) Quality assurance framework of the european statistical system. https://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646, accessed 07.05.2020

Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery in databases. AI Magazine 17(3):37–37

FDA (2019) https://www.fda.gov/media/122535/download, accessed 13.05.2020

Flake JK, Fried EI (2020) Measurement schmeasurement: questionable measurement practices and how to avoid them. Adv Methods Practices Psychol Sci 3(4):456–465. https://doi.org/10.1177/2515245920952393

Forbes (2018) https://www.forbes.com/sites/bernardmarr/2018/03/05/heres-why-data-is-not-the-new-oil/#45b487143aa9, accessed 27.04.2020

Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55(1):119–139. https://doi.org/10.1006/jcss.1997.1504

Friede T, Nicholas R, Stallard N, Todd S, Parsons N, Valdés-Márquez E, Chataway J (2010) Refinement of the clinical scenario evaluation framework for assessment of competing development strategies with an application to multiple sclerosis. Drug Inf J: DIJ/Drug Inf Assoc 44(6):713–718

Friedrich S, Groß S, König IR, Engelhardt S, Bahls M, Heinz J, Huber C, Kaderali L, Kelm M, Leha A, Rühl J, Schaller J, Scherer C, Vollmer M, Seidler T, Friede T (2021) Applications of AI/ML approaches in cardiovascular medicine: A systematic review with recommendations. European Heart Journal - Digital Health. https://doi.org/10.1093/ehjdh/ztab054

Gabler S, Häder S (2018) Repräsentativität: Versuch einer Begriffsbestimmung. In: Telefonumfragen in Deutschland, Springer Fachmedien Wiesbaden, pp 81–112, https://doi.org/10.1007/978-3-658-23950-3_5, https://doi.org/10.1007%2F978-3-658-23950-3_5

Gal Y, Ghahramani Z (2016) Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: Balcan MF, Weinberger KQ (eds) Proceedings of The 33rd international conference on machine learning, PMLR, New York, New York, USA, Proceedings of Machine Learning Research, vol 48, pp 1050–1059

Garnelo M, Rosenbaum D, Maddison CJ, Ramalho T, Saxton D, Shanahan M, Teh YW, Rezende DJ, Eslami S (2018) Conditional neural processes. arXiv preprint arXiv:1807.01613

Gaye A, Marcon Y, Isaeva J, LaFlamme P, Turner A, Jones EM, Minion J, Boyd AW, Newby CJ, Nuotio ML (2014) DataSHIELD: taking the analysis to the data, not the data to the analysis. Int J Epidemiol 43(6):1929–1944

Greenland S, Robins JM (1986) Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol 15(3):413–419

Greenland S, Robins JM, Pearl J (1999) Confounding and collapsibility in causal inference. Stat Sci 14(1):29–46

Gudivada VN, Baeza-Yates R, Raghavan VV (2015) Big data: Promises and problems. Computer 48(3):20–23. https://doi.org/10.1109/MC.2015.62

Györfi L, Kohler M, Krzyzak A, Walk H (2002) A distribution-free theory of nonparametric regression. Springer, New York. https://doi.org/10.1007/b97848

Haibe-Kains B, Adam GA, Hosny A, Khodakarami F, Waldron L, Wang B, McIntosh C, Goldenberg A, Kundaje A, Greene CS et al (2020) Transparency and reproducibility in artificial intelligence. Nature 586(7829):E14–E16

Hamburg MA, Collins FS (2010) The path to personalized medicine. N Engl J Med 363(4):301–304

Heckman JJ (2001) Micro data, heterogeneity, and the evaluation of public policy: nobel lecture. J Political Econ 109(4):673–748. https://doi.org/10.1086/322086

Heinze G, Wallisch C, Dunkler D (2018) Variable selection-a review and recommendations for the practicing statistician. Biomet J 60(3):431–449

Higgins JP, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Savović J, Schulz KF, Weeks L, Sterne JA (2011) The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. Bmj 343:d5928

Hilberg W (1995) Karl Steinbuch, ein zu Unrecht vergessener Pionier der künstlichen neuronalen Systeme. Frequenz 49(1–2):28–36

Hill AB (1965) The environment and disease: association or causation? Proc Royal Soc Med 58(5):295–300

Huang GB, Zhu QY, Siew CK (2006) Extreme learning machine: Theory and applications. Neurocomputing 70(1–3):489–501. https://doi.org/10.1016/j.neucom.2005.12.126

Juang BH, Rabiner LR (1991) Hidden markov models for speech recognition. Technometrics 33(3):251–272. https://doi.org/10.1080/00401706.1991.10484833

Kamilaris A, Prenafeta-Boldú FX (2018) Deep learning in agriculture: A survey. Comput Electron Agric 147:70–90

Karlin S, Studden WJ (1966) Optimal experimental designs. Ann Math Stat 37(4):783–815

Karr AF, Sanil AP, Banks DL (2006) Data quality: A statistical perspective. Stat Methodol 3(2):137–173

Kearns MJ, Vazirani U (1994) An introduction to computational learning theory. The MIT Press, Cambridge, MA. https://doi.org/10.7551/mitpress/3897.001.0001

Kelley HJ (1960) Gradient theory of optimal flight paths. ARS J 30(10):947–954. https://doi.org/10.2514/8.5282

Keogh RH, Shaw PA, Gustafson P, Carroll RJ, Deffner V, Dodd KW, Küchenhoff H, Tooze JA, Wallace MP, Kipnis V et al (2020) Stratos guidance document on measurement error and misclassification of variables in observational epidemiology: part 1–basic theory and simple methods of adjustment. Stat Med 39(16):2197–2231

Koch C (2016) How the computer beat the go player. Sci Am Mind 27(4):20–23. https://doi.org/10.1038/scientificamericanmind0716-20

Kohavi R, Tang D, Xu Y, Hemkens LG, Ioannidis JPA (2020) Online randomized controlled experiments at scale: lessons and extensions to medicine. Trials 21(1), https://doi.org/10.1186/s13063-020-4084-y, https://doi.org/10.1186%2Fs13063-020-4084-y

Kozielski M, Doetsch P, Ney H (2013) Improvements in RWTH's System for Off-Line Handwriting Recognition. In: 2013 12th international conference on document analysis and recognition, IEEE, https://doi.org/10.1109/icdar.2013.190, https://doi.org/10.1109%2Ficdar.2013.190

Kruskal W, Mosteller F (1979a) Representative sampling, I: non-scientific literature. International Statistical Review/Revue Internationale de Statistique pp 13–24

Kruskal W, Mosteller F (1979b) Representative sampling. Scientific literature, excluding statistics. International Statistical Review/Revue Internationale de Statistique, II, pp 111–127

Kruskal W, Mosteller F (1979c) Representative sampling. The current statistical literature. International Statistical Review/Revue Internationale de Statistique, III, pp 245–265

Kruskal W, Mosteller F (1980) Representative sampling, IV: The history of the concept in statistics, 1895-1939. International Statistical Review/Revue Internationale de Statistique pp 169–195

Van der Laan MJ, Rose S (2011) Targeted learning: causal inference for observational and experimental data. Springer, New York

Langford J (2005) Tutorial on practical prediction theory for classification. J Mach Learn Res 6:273–306

Lazer D, Kennedy R, King G, Vespignani A (2014) The parable of Google Flu: traps in big data analysis. Science 343(6176):1203–1205

Levine S, Pastor P, Krizhevsky A, Ibarz J, Quillen D (2018) Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. Int J Robot Res 37(4–5):421–436

Lin EJD, Hefner JL, Zeng X, Moosavinasab S, Huber T, Klima J, Liu C, Lin SM (2019) A deep learning model for pediatric patient risk stratification. Am J Managed Care 25(10):e310–e315

Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK (2020) Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. British Med J 370:m3164

Makel MC, Plucker JA, Hegarty B (2012) Replications in psychology research: How often do they really occur? Perspec Psychol Sci 7(6):537–542

McCracken MW, Ng S (2016) FRED-MD: a monthly database for macroeconomic research. J Business Econ Stat 34(4):574–589. https://doi.org/10.1080/07350015.2015.1086655

MedTechIntelligence (2018) https://www.medtechintelligence.com/news_article/apple-watch-4-gets-fda-clearance/, accessed 13.05.2020

Meinshausen N, Bühlmann P (2010) Stability selection. J Royal Stat Soc: Ser B (Statistical Methodology) 72(4):417–473

Meng XL (2018) Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. Ann Appl Stat 12(2):685–726

Meng XL, Xie X (2014) I got more data, my model is more refined, but my estimator is getting worse! Am I just dumb? Econom Rev 33(1–4):218–250

Miller T (2019) Explanation in artificial intelligence: Insights from the social sciences. Artif Intell 267:1–38

(2014) Handbook of missing data methodology. Chapman and Hall/CRC, Boca Raton, FL,. https://doi.org/10.1201/b17622

Molnar C (2019) Interpretable machine learning. https://christophm.github.io/interpretable-ml-book/. Accessed 29 July 2020

Moor J (2006) The Dartmouth College artificial intelligence conference: The next fifty years. AI Magazine 27(4):87–87

Morris TP, White IR, Crowther MJ (2019) Using simulation studies to evaluate statistical methods. Stat Med 38(11):2074–2102

New York A (2018) https://www.nytimes.com/2018/12/18/technology/facebook-privacy.html, accessed 27.04.2020

Ng S (2018) Opportunities and challenges: lessons from analyzing terabytes of scanner data. In: Honore B, Pakes A, Piazzesi M, Samuelson L (eds) Advances in economics and econometrics, Cambridge University Press, pp 1–34, https://doi.org/10.1017/9781108227223.001, https://doi.org/10.1017%2F9781108227223.001

Ntoutsi E, Fafalios P, Gadiraju U, Iosifidis V, Nejdl W, Vidal ME, Ruggieri S, Turini F, Papadopoulos S, Krasanakis E et al (2020) Bias in data-driven artificial intelligence systems. An introductory survey. Wiley Interdisciplin Rev: Data Mining Knowl Discovery 10(3):e1356

Nuffield Foundation (2019) Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research. https://www.nuffieldfoundation.org/sites/default/files/files/Ethical-and-Societal-Implications-of-Data-and-AI-report-Nuffield-Foundat.pdf, accessed 27.04.2021

Osband I, Blundell C, Pritzel A, Van Roy B (2016) Deep exploration via bootstrapped DQN. In: Advances in neural information processing systems, pp 4026–4034

Pashler H, Wagenmakers EJ (2012) Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? Perspect Psychol Sci 7(6):528–530

Pearl J (1988) Probabilistic reasoning in intelligent systems: Networks of plausible inference. Morgan Kaufmann Publisher Inc, San Francisco, CA

Pearl J (1993) Aspects of graphical models connected with causality. In: Proceedings of the 49th session of the international statistical science institute

Pearl J (2009) Causality. Cambridge University Press, New York

Pearl J (2010) The foundations of causal inference. Sociol Methodol 40(1):75–149

Pearl J (2018) Theoretical impediments to machine learning with seven sparks from the causal revolution. arXiv preprint arXiv:18010.4016v1

Peltola T (2018) Local interpretable model-agnostic explanations of bayesian predictive models via Kullback–Leibler projections. arXiv preprint arXiv:18100.2678v1

Pepe MS (2003) The statistical evaluation of medical tests for classification and prediction. Oxford University Press, New York

Perez MV, Mahaffey KW, Hedlin H, Rumsfeld JS, Garcia A, Ferris T, Balasubramanian V, Russo AM, Rajmane A, Cheung L et al (2019) Large-scale assessment of a smartwatch to identify atrial fibrillation. N Engl J Med 381(20):1909–1917

Porta M (ed) (2016) A Dictionary of Epidemiology, 6th edn. Oxford University Press, New York

Ramosaj B, Pauly M (2019a) Consistent estimation of residual variance with random forest Out-Of-Bag errors. Stat Probab Lett 151:49–57

Ramosaj B, Pauly M (2019b) Predicting missing values: a comparative study on non-parametric approaches for imputation. Comput Stat 34(4):1741–1764

Ramosaj B, Amro L, Pauly M (2020) A cautionary tale on using imputation methods for inference in matched pairs design. Bioinformatics 36(10):3099–3106

Ribeiro M, Singh S, Guestrin C (2016a) "Why Should I Trust You?": Explaining the predictions of any classifier. In: Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: Demonstrations, Association for Computational Linguistics, https://doi.org/10.18653/v1/n16-3020, https://doi.org/10.18653%2Fv1%2Fn16-3020

Ribeiro MT, Singh S, Guestrin C (2016b) Model-agnostic interpretability of machine learning. arXiv preprint arXiv:16060.5386v1

Richter J, Madjar K, Rahnenführer J (2019) Model-based optimization of subgroup weights for survival analysis. Bioinformatics 35(14):484–491

Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ (2020) Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. British Med J 370:m3210

Robins JM, Hernán MÁ, Brumback B (2000) Marginal structural models and causal inference in epidemiology. Epidemiology 11(5):550–560. https://doi.org/10.1097/00001648-200009000-00011

Roe BE, Just DR (2009) Internal and external validity in economics research: tradeoffs between experiments, field experiments, natural experiments, and field data. Am J Agricult Econom 91(5):1266–1271. https://doi.org/10.1111/j.1467-8276.2009.01295.x

Rosenbaum P (2002) Observational studies. In: Springer Series in Statistics, Springer New York, pp 1–17, https://doi.org/10.1007/978-1-4757-3692-2_1, https://doi.org/10.1007%2F978-1-4757-3692-2_1

Rosenbaum P (2010) Design of observational studies. Springer, New York. https://doi.org/10.1007/978-1-4419-1213-8

Rosenbaum P (2017) Observation and experiment. Harvard University Press, Cambridge, MA. https://doi.org/10.4159/9780674982697

Rosenblatt F (1958) The perceptron: A probabilistic model for information storage and organization in the brain. Psychol Rev 65(6):386–408. https://doi.org/10.1037/h0042519

Ross A, Lage I, Doshi-Velez F (2017) The neural lasso: Local linear sparsity for interpretable explanations. In: Workshop on transparent and interpretable machine learning in safety critical environments, 31st conference on neural information processing systems, Long Beach, CA

Röver C, Friede T (2020) Dynamically borrowing strength from another study through shrinkage estimation. Stat Methods Med Res 29:293–308

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 66(5):688

Rubin DB (1976) Inference and missing data. Biometrika 63(3):581–592

Rubin DB (2006) Matched sampling for causal effects. Cambridge University Press, Cambridge, MA. https://doi.org/10.1017/cbo9780511810725

Rubin DB (2008) For objective causal inference, design trumps analysis. Ann Appl Stat 2(3):808–840

Sauerbrei W, Perperoglou A, Schmid M, Abrahamowicz M, Becher H, Binder H, Dunkler D, Harrell FE, Royston P, Heinze G, others for TG2 of the STRATOS initiative (2020) State of the art in selection of variables and functional forms in multivariable analysis - outstanding issues. Diagnostic Prognostic Res 4:1–18

Schmidhuber J (2015) Deep learning in neural networks: An overview. Neural Netw 61:85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Schork NJ (2019) Artificial intelligence and personalized medicine. In: Von Hoff D, Han H (eds) Precision medicine in cancer therapy, cancer treatment and research. Springer, Cham

Scornet E, Biau G, Vert JP (2015) Consistency of random forests. Ann Stat 43(4):1716–1741

Seaman SR, White IR (2013) Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res 22(3):278–295

Searle J (1980) Minds, Brains and Programs. Behavioral Brain Sci 3(3):417–457

Shadish WR, Cook TD, Campbell DT (2002) Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin, Boston

Shaw PA, Gustafson P, Carroll RJ, Deffner V, Dodd KW, Keogh RH, Kipnis V, Tooze JA, Wallace MP, Küchenhoff H et al (2020) Stratos guidance document on measurement error and misclassification of variables in observational epidemiology: Part 2–more complex methods of adjustment and advanced topics. Stat Med 39(16):2232–2263

Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guez A, Lanctot M, Sifre L, Kumaran D, Graepel T, Lillicrap T, Simonyan K, Hassabis D (2018) A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science 362(6419):1140–1144. https://doi.org/10.1126/science.aar6404

Simon HA (1983) Why should machines learn? In: Michalski RS, Carbonell JG, Mitchell TM (eds) Machine learning. Morgan Kaufmann, San Francisco, CA, pp 25–37

Simons DJ, Shoda Y, Lindsay DS (2017) Constraints on generality (COG): A proposed addition to all empirical papers. Perspect Psychol Sci 12(6):1123–1128

Simpson EH (1951) The interpretation of interaction in contingency tables. J Roy Stat Soc: Ser B (Methodol) 13(2):238–241

Solomonoff RJ (1985) The time scale of artificial intelligence: Reflections on social effects. Human Syst Manag 5(2):149–153

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958

Staddon J (2017) Scientific method: how science works, fails to work, and pretends to work. Taylor & Francis Group, New York

Steinbuch K (1961) Die Lernmatrix. Kybernetik 1(1):36–45

Sutton RS, Barto AG (2018) Reinforcement learning: An introduction. MIT press, Cambridge, MA

Teichmann M, Weber M, Zoellner M, Cipolla R, Urtasun R (2018) Multinet: Real-time joint semantic reasoning for autonomous driving. In: 2018 IEEE intelligent vehicles symposium (IV), IEEE, pp 1013–1020

The Economist (2017) https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data, accessed 27.04.2020

Theodorou V, Abelló A, Thiele M, Lehner W (2017) Frequent patterns in ETL workflows: An empirical approach. Data Knowl Eng 112:1–16. https://doi.org/10.1016/j.datak.2017.08.004

Thurow M, Dumpert F, Ramosaj B, Pauly M (2021) Goodness (of fit) of imputation accuracy: The Good-Impact analysis. arXiv preprint arXiv:2101.07532

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Royal Stat Soc Ser B Stat Methodol 58(1):267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tibshirani R (1997) The LASSO method for variable selection in the Cox model. Stat Med 16(4):385–395. https://doi.org/10.1002/(sici)1097-0258(19970228)16:4h385::aid-sim380i3.0.co;2-3

Tukey JW (1962) The future of data analysis. Ann Math Stat 33(1):1–67

UNECE (2020) Machine learning for official statistics – HLG-MOS machine learning project. https://statswiki.unece.org/display/ML/HLG-MOS+Machine+Learning+Project

Valiant LG (1984) A theory of the learnable. Commun ACM 27(11):1134–1142. https://doi.org/10.1145/1968.1972

Valiant LG (2013) Probably approximately correct: nature's algorithms for learning and prospering in a complex world. Choice Rev Online 51(05):51-2716–51-2716. https://doi.org/10.5860/choice.51-2716

Van Buuren S (2018) Flexible imputation of missing data. CRC Press, Boca Raton, FL

Vapnik V (1998) Statistical learning theory. Wiley, New York

Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. J Am Stat Assoc 113(523):1228–1242

Wager S, Wang S, Liang PS (2013) Dropout training as adaptive regularization. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (eds) Advances in neural information processing systems, vol 26. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2013/file/38db3aed920cf82ab059bfccbd02be6a-Paper.pdf

Wager S, Hastie T, Efron B (2014) Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. J Mach Learn Res 15(1):1625–1651

Warner B, Misra M (1996) Understanding neural networks as statistical tools. Am Stat 50(4):284–293

Weihs C, Ickstadt K (2018) Data science: the impact of statistics. Int J Data Sci Anal 6(3):189–194

Wickramasuriya SL, Athanasopoulos G, Hyndman RJ (2019) Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. J Am Stat Assoc 114(526):804–819

Wikipedia (2020) https://en.wikipedia.org/wiki/Simpson%27s_paradox#/media/File:Simpson's_paradox_continuous.svg. Accessed 28 July 2020

Wiredcom (2019) https://www.wired.com/story/ubers-self-driving-car-didnt-know-pedestrians-could-jaywalk/. 13 May 2020

Wolf MJ, Miller K, Grodzinsky FS (2017) Why we should have seen that coming: comments on Microsoft's tay experiment, and wider implications. ACM SIGCAS Comput Soc 47(3):54–64

Zaremba W, Sutskever I, Vinyals O (2014) Recurrent Neural Network Regularization. arXiv preprint arXiv:1409.2329v5

Zhu J, Chen J, Hu W, Zhang B (2017) Big Learning with Bayesian methods. National Sci Rev 4(4):627–651. https://doi.org/10.1093/nsr/nwx044

Affiliations

Sarah Friedrich[1] 📍 · **Gerd Antes**[2] · **Sigrid Behr**[3] · **Harald Binder**[2] ·
Werner Brannath[4] · **Florian Dumpert**[5] · **Katja Ickstadt**[6] · **Hans A. Kestler**[7] ·
Johannes Lederer[8] · **Heinz Leitgöb**[9] · **Markus Pauly**[6] · **Ansgar Steland**[10] ·
Adalbert Wilhelm[11] · **Tim Friede**[1] 📍

[1]   Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany

[2]   Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany

[3]   Novartis Pharma AG, Basel, Switzerland

[4]   Competence Center for Clinical Trials and Institute of Statistics, University Bremen, Bremen, Germany

[5]   Federal Statistical Office of Germany, Wiesbaden, Germany

[6]   Department of Statistics, TU Dortmund University, Dortmund, Germany

[7]   Institute of Medical Systems Biology, Ulm University, Ulm, Germany

[8]   Faculty of Mathematics, Ruhr-Universität Bochum, Bochum, Germany

[9]   Department of Sociology, University of Eichstätt-Ingolstadt, Ingolstadt, Germany

[10]   RWTH Aachen University, Aachen, Germany

[11]   Jacobs University Bremen, Bremen, Germany