Journal of Methods and Measurement in the Social Sciences,
Vol. 13, No.x , xx-xx, 2022

Comparing human coding to two natural language processing algorithms in aspirations of people affected by Duchenne Muscular Dystrophy

Carolyn E. Schwartz
DeltaQuest Foundation
Tufts University Medical School

Roland B. Stark
DeltaQuest Foundation

Elijah Biletch
DeltaQuest Foundation

Richard B.B. Stuart
DeltaQuest Foundation

Yuelin Li
Memorial Sloan Kettering Cancer Center

Qualitative methods can enhance our understanding of constructs that have not been well portrayed and enable nuanced depiction of experience from study participants who have not been broadly studied. However, qualitative data require time and effort to train raters to achieve validity and reliability. This study compares recent advances in Natural Language Processing (NLP) models with human coding. This web-based study (N=1,253; 3,046 free-text entries, averaging 64 characters per entry) included people with Duchenne Muscular Dystrophy (DMD), their siblings, and a representative comparison group. Human raters (n=6) were trained over multiple sessions in content analysis as per a comprehensive codebook. Three prompts addressed distinct aspects of participants' aspirations. Unsupervised NLP was implemented using Latent Dirichlet Allocation (LDA), which extracts latent topics across all the free-text entries. Supervised NLP was done using a Bidirectional Encoder Representations from Transformers (BERT) model, which requires training the algorithm to recognize relevant human-coded themes across free-text entries. We compared the human-, LDA-, and BERT-coded themes. Study sample contained 286 people with DMD, 355 DMD siblings, and 997 comparison participants, age 8-69. Human coders generated 95 codes across the three prompts and had an average inter-rater reliability (Fleiss's kappa) of 0.77, with minimal rater-effect (pseudo R^2=4%). Compared to human coders, LDA does not yield easily interpretable themes. BERT correctly classified only 61-70% of the validation set. LDA and BERT required technical expertise to program and took approximately 1.15 minutes per open-text entry, compared to 1.18 minutes for human raters including training time. LDA and BERT provide potentially viable approaches to analyzing large-scale qualitative data, but both have limitations. When text entries are short, LDA yields latent topics that are hard to interpret. BERT accurately identified only about two thirds of new statements. Humans provided reliable and cost-effective coding in the web-based context. The upfront training enables BERT to process enormous quantities of text data in future work, which should examine NLP's predictive accuracy given different quantities of training data.

Key words: natural language processing, qualitative data, human, efficiency

While qualitative data collection is often used in the development of theory or conceptual models for new measures (Cappelleri et al., 2013; Ferrans, 2005), many qualitative studies utilize small sample sizes (Schwartz & Revicki, 2012), perhaps related to different logical, theoretical, and epistemological differences from quantitative research (Trotter II, 2012). There are, however, increasingly low-effort ways to collect qualitative data due to online survey engines, social media platforms, and other ways that people are asked to provide input in their clinical care. These developments dovetail with the growing interest in patient-centered measurement and care (Barry & Edgman-Levitan, 2012; Kebede, 2016), providing further motivation for expanding the feasibility and use of qualitative data in outcome research. Moreover, advances in the capability of Natural Language Processing (NLP) algorithms over the past decade have expanded their applications in medical and social science research (Agaronnik, Lindvall, El-Jawahri, He, & Iezzoni, 2020; Parker, 2020; Skaljic et al., 2019).

Early NLP algorithms extracted themes by tallying word frequencies across responses. One of the most widely used instances of this basic type of NLP software is Linguistic Inquiry and Word Count (LIWC) (Pennebaker, Francis, & Booth, 2001). LIWC compares words to a pre-determined dictionary file of various linguistic and psychological categories, allowing researchers to observe categorical associations between linguistic patterns and psychological state (Dönges, 2009; Pennebaker et al., 2001; Receptiviti, 2021). A 2011 study, for example, used LIWC to analyze transcripts of potential romantic partners on four-minute speed dates to measure how closely their speech matched in order to predict whether the couples would stay together after the first date (Ireland et al., 2011).

The major limitation of word-count algorithms such as LIWC is the requirement for investigators to predict words and categories relevant to the research topic for use in the algorithm's "dictionary" in order to be counted for analysis. Topic modelling algorithms from the mid-2000's such as Latent Dirichlet Analysis (LDA) and Hierarchical Latent Tree Analysis solved this problem by generating lists of abstract topics from text without the need for a "warmup" or "training" data set. Though researchers may use topics to extract information such as patient priorities and goals, the topics produced by LDA and Hierarchical Latent Tree Analysis are often unpolished and may not be relevant to the research question (Atkinson, 2019; Li, Rapkin, Atkinson, Schofield & Bochner, 2019).

Technology companies such as Facebook, Google, and OpenAI have recently developed deep learning neural network tools and made many of them open source and free to download. In this article, we applied a Bidirectional Encoder Representations from Transformers (BERT) model to classify free-text goal statements to themes (Devlin, Chang, Lee, & Toutanova, 2018). Transformers first appeared in 2017 (Vaswani et al., 2017), when engineers at Google published a paper to address challenges in processing word sequences. For example, the two phrases “live to eat” and “eat to live” are semantic opposites due to how the words are ordered from left to right. Transformers use an attention-based structure to retain a memory of word sequences in hidden layers. This attention mechanism overcomes the limitations of LDA, which has no built-in mechanisms to distinguish word sequences (except in n-grams, sequences of words treated as unique entities, but it is a flawed approach). Word order is also ignored in machine-learning techniques such as naive bayes, Support Vector Machines (SVM), and random forest (Reyes, 2019).

In late 2018, Google released BERT (Devlin et al., 2018) which added enhancements to the attention mechanisms of Transformers. Given that generally the larger the quantities of data used to train a neural network, the more the predictive power, the immense network of data available to research groups at the technology companies have allowed for the development of these NLP techniques that more accurately assess nuanced contexts and motives in individuals’ writing and speech (Mikolov, Chen, Corrado, & Dean, 2013; Tenney, Das, & Pavlick, 2019). Notably, this improved accuracy has allowed for the development of NLP systems capable of deriving clinical decisions based on automated electronic medical record analysis (Chen, Zafar, Galperin-Aizenberg, & Cook, 2018; Gonzalez-Hernandez, Sarker, O'Connor, & Savova, 2017). The primary use of NLP in social science and medical research, however, is to supersede the use of humans in assigning topic “codes” to open-text survey responses, interviews, and social media posts (Guetterman et al., 2018; Leeson, Resnick, Alexander, & Rovers, 2019).

Early articles comparing NLP to human coding were optimistic about its potential; Andrew Perrin postulated that NLP could expand the scope of qualitative studies by eliminating the need to pay and train coders and could potentially even eliminate issues regarding inter-rater reliability, though computer processing power at the time did not yet allow NLP to outpace human coders and thus limited its applications (Perrin, 2001). Accordingly, newer computing technologies have yielded promising results in certain fields; for instance, LDA topic modeling analysis of open-ended survey questions can allow for thematic information outside of a predefined coding rubric to be detected in survey responses, which serves to augment, rather than replace, the manual coding of data (Finch, Hernández Finch, McIntosh, & Braun, 2018).

Counseling psychology studies comparing NLP analysis to human coding of counselor-client conversations/motivational interviews have also found evidence that NLP techniques may be able to accurately apply a behavioral coding system on a large body of unstructured text. This may save significant time and money over a manual approach, which can range on average from 90 to 120 minutes per 20 minute interview segment, not including the 40 or so hours required for coder training (Can et al., 2016; Moyers, Martin, Catley, Harris, & Ahluwalia, 2003). Some non-topic models lagged behind human reliability when coding certain highly contextual statements in motivational interviews; in one example, the Discrete Sentence Feature (DSF) and Recursive Neural Network (RNN) models struggled with coding isolated sentences discussing substance use. Those sentences could either be coded as favoring change in the client's habits or as the opposite (favoring maintenance of current habits), depending on subtle context clues from the preceding conversation, which human raters found easier to discern (Tanana, Hallgren, Imel, Atkins, & Srikumar, 2016). A number of these studies express optimism about the potential speed advantage of NLP over human coders.

Baumer and colleagues compared LDA and human coding by grounded theory (Baumer, Mimno, Guha, Quan, & Gay, 2017). BERT was not yet available at the time of their application. They analyzed free-text data on reasons why individuals returned to social media after a brief (up to 99 days) and voluntary absence. Baumer et al. (2017) report good agreement, that LDA extracts themes that generally reflect the same content as the human-extracted themes.

In response to a dearth of literature (Raffel et al., 2019), the present study directly compared human coding to two NLP methods: the un-supervised LDA and the supervised BERT. One of the co-authors (YL) has previously applied LDA to summarize cancer patients' free-text goal statements as they undergo bladder cancer surgery (Landis & Koch, 1977). Thus, the main rationale for these two specific NLP methods is to go beyond LDA to capitalize on the latest NLP analytics.

Methods

Sample and Procedure

This secondary analysis utilized data from a study of Duchenne Muscular Dystrophy (DMD) patients, their siblings, and a comparison-group. The sample and methods are fully described in the two primary papers from this project (Schwartz, Biletch, Stuart, & Rapkin, 2022a, 2022b) and will only briefly summarized herein. These primary papers examine differences in aspirations for patients versus comparison participants, and siblings versus comparison participants. Accordingly, for the purpose of the present work, data were combined across groups, although demographic characteristics will be described by group. Eligible participants were age 8 or older and able to complete an online questionnaire.

The web-based survey was administered October through December 2020 through the Health Insurance Portability and Accountability Act of

1996 (HIPAA)-compliant, secure Alchemer engine (www.alchemer.com). Participants were paid honoraria to compensate them for their time. The protocol was reviewed and approved by the New England Independent Review Board (NEIRB #20203038), and all participants provided informed consent before beginning the survey.

Measures

Life aspirations was measured using the following open-ended prompts: (1) Three Wishes (Nereo & Hinton, 2003), in which participants were asked, "If you could make three wishes, any three wishes in the whole world, what would they be?"; (2) Goals: "What are the main things you want to accomplish?"; (3) Quality of Life (QOL) Definition: "In a sentence, what does the phrase "Quality of Life" mean to you at this time?" The latter two are part of the QOL Appraisal Profile$_{V1}$(QOLAP) (Rapkin & Schwartz, 2004).

Demographic Characteristics included year of birth, gender, and whether anyone in the household was or had been infected with the novel coronavirus-2019, and whether they received help completing the survey (all participants). Teens and adults were asked about comorbidities from a list selected on the basis of documented higher prevalence in people with DMD (Ciafaloni et al., 2009; Pane et al., 2012). Adult participants were asked about race, ethnicity, education, marital status, weight, height, with whom the person lives, difficulty paying bills, and employment status.

Statistical Analysis

Coding open-text data. The open-ended data were coded by six trained raters (EB, RBB, AD, JBL, EK, MCF), according to a standardized protocol and comprehensive codebook originally derived from an extensive sorting procedure (Li & Rapkin, 2009). [The interested reader can contact the corresponding author for the QOLAP coding manual which describes the theme definitions in detail.]

Themes were coded as "0" if they were not reflected in the individual's written text response, and "1" if they were reflected there. As the goal-delineation themes were originally developed with a Human Immunodeficiency Virus sample (Li & Rapkin, 2009), which generally has different sociodemographic characteristics than the current study sample, some themes were not as prevalent among the present sample. For example, themes related to drug and alcohol, immigration, and racism were prevalent among the Human Immunodeficiency Virus sample, but were not found at all in the current study sample. Themes were added as needed, resulting in a set of 40 themes for the Wishes and Goals prompts and 17 for the QOL Definition prompt. For each prompt, a theme of "no direct answer" was used if the respondent did not provide an answer or answering a different question than was asked. For example, in response to the question "What are the main things you want to accomplish?" exemplary No-Direct-Answer responses "seems rather great" or "nothing idk lol."

Each text entry could be coded for as many themes as there. Thus, one goal could elicit one theme or more than one depending on how the individual worded it. For example, one individual's Accomplish goal was "My bills paid, my family healthy and happy, and family go to church". It was coded as reflecting family welfare, financial concerns, health issues, mental health/mood state, and religious/spiritual concerns. In contrast, another individual's Accomplish goal was "Move to a different state," Which was coded only as living situation. In this method of working with the aspirations data, we assumed that the relevant factor was the themes, not the different wishes, goals, or QOL definitions themselves.

Training took place in two multi-hour sessions to understand the protocol and to utilize fully the codebook where themes were described fully and exemplified. Raters coded an initial set of ten participants' data (all prompts), followed by a discussion of differences across raters. They then coded the next ten participants' data (all prompts), and comparison and discussion revealed almost no differences across raters. Raters then coded data from 40 more responses (all prompts), from which inter-rater reliability was computed in two ways on the 240 test responses (6 raters * 40 participant entries).

Inter-Rater Reliability. Fleiss's kappa (Fleiss, 1971) assessed degree of agreement over and above what would be expected by chance. This variant on the more familiar Cohen's kappa (Cohen, 1960) is used in cases of more than two raters. While there are no generally accepted rules of thumb for a desirable level of either form of kappa, some healthcare researchers have proposed values from 0.41-0.60 as "moderate," 0.61-0.80 as "good," and 0.81-1.00 as "very good."(Altman, 1999; Landis & Koch, 1977).

Logistic regression assessed level of agreement among raters, with each of 240 "0" or "1" values regressed on the Rater variable, with its six rater-categories. High inter-rater reliability (IRR) for any given theme would be indicated by a nonsignificant rater effect, and one that explained a low fraction of the variance in ratings (e.g., a pseudo-R-squared in the low single digits).

NLP Methods Tested

Two NLP methods were tested in this study: LDA and BERT. The main difference between these methods is that LDA is unsupervised, and BERT is supervised machine learning, in the sense that LDA is able to extract topics without human intervention while BERT (in text classification specifically) requires that topics be previously established. A crude but useful analogy may be that LDA behaves more like Exploratory Factor

Analysis, where the underlying factors are unknown, while BERT behaves more like Confirmatory Factor Analysis, where those factors are specified in advance.

LDA. The LDA analytic plan was similar to the one described in detail in a previous article on patients' free-text goal statements as they undergo bladder cancer surgery.(Li, et al., 2019) Separate LDA analyses were conducted for responses to each of three prompts: Wishes, QOL Definitions, and Goals. We followed the commonly-used steps in preprocessing (e.g., plotting 'word clouds', setting 'stop words' aside, and adding two-consecutive-word phrases as 'bigrams' for contextual information). We then determined the best number of topics as specified by LDA and fitted the final LDA model for each analysis. The LDA computation was primarily done by the scikit-learn tools written in the Python programming language (Pedregosa et al., 2011). The number of topics per analysis was evaluated by the R package ldatuning (Nikita, 2016) and the four supported metrics (Arun, Suresh, Veni Madhavan, & Murthy, 2010; Cao, Xia, Li, Zhang, & Tang, 2009; Deveaud, SanJuan, & Bellot, 2014; Griffiths & Steyvers, 2004), using all available text entries. The LDA analysis, unlike that for BERT, involved no evaluation of accuracy, as in use of a training set versus validation set.

Model selection was done using the four metrics provided in the ldatuning package (Nikita, 2016) to estimate the desired number of topics. Both the Arun et al. (2010) and Cao et al. (2009) metrics are akin to the scree plot in an exploratory factor analysis, where the location of the elbow indicates the desired number of topics. The Griffiths and Steyvers' (2004) and the Deveaud (2014) metric are based on the fit between words within topics, where the location of a plateau reflects the desired number of topics.

All subsequent analyses were fixed at this number of topics to make a consistent and streamlined presentation, including separate LDA models for patients, siblings, and comparison-group participants.

Bidirectional Encoder Representations from Transformers (BERT). BERT is widely viewed as a state-of-the-art, supervised deep-learning neural network. It was developed by scientists at Google (Devlin, Chang, Lee, & Toutanova, 2019) to address enduring challenges in NLP. Transformers such as BERT use an attention-based structure to retain a memory of word sequences in hidden layers of a neural network such that the network registers or "intuitively understands" their opposite semantic meaning. This property overcomes certain limitations of LDA, which has no built-in mechanisms to distinguish word sequences (except in n-grams, such as the bigram in the current LDA approach, an unsatisfactory workaround nevertheless).

To classify text using BERT. we used a publicly accessible, off-the-shelf machine-learning tool called the "huggingface transformers" (Hugging Face, 2021). The specific tool we used was the DistillBERT tool within the huggingface transformers library. This a scaled-down version of the full BERT was designed to work more quickly due to fewer layers and hidden nodes. It is one of several alternative algorithms derived from the full BERT technology (see ("List of alternative algorithms derived from the full BERT technology,")). DistillBERT is what is known as a *pre-trained* model, in which technology companies have already trained it using the enormous amounts of unannotated text on the internet so that it learns a general-purpose language representation model (Devlin & Chang, 2018). After pre-training, an analyst can then fine-tune DistillBERT for specific tasks. Henceforth, for simplicity and readability we use the more generic term BERT to represent DistillBERT.

From a user's perspective, an application of BERT is divided into two components, known in the literature as *pre-training* and *fine-tuning*. This two-step approach is at the core of the concept of Transfer Learning(Vaswani et al., 2017). Once pre-trained, BERT and its variants can be reused for many downstream machine-learning tasks, including the current text classification. There are many pre-trained libraries available for download, for tasks such as next-sentence prediction (e.g., instant autocomplete suggestions in a search engine), named-entity recognition (e.g., a trained network knows that the Empire State Building is near Manhattan), and language translation (e.g., English to French).

The fine-tuning in this study proceeded as follows. Wishes, goals and definitions were analyzed separately. For example, the 1,613 entries of wishes were randomly divided into the training set (n=399), the validation set (n=76 for tuning configuration parameters), and a blinded test set (remaining n=1,214 that BERT had never encountered previously and blinded to the analyst who trained BERT).

Configuration Parameters for BERT. The training set entries were entered into BERT as the predictors and the corresponding human-coded categories were the target outcomes. Learning was achieved by optimizing network connections by the Adaptive Moment Estimation algorithm (Kingma & Ba, 2014). It is known that optimized network configurations are affected by hyperparameters such as the learning rate (the rate with which model weights are updated in response to the estimated error, where a learning rate too small may run slowly but a learning rate too large may lead to suboptimal weights), batch size (number of samples that are passed to the network at once, where smaller batch size facilitates learning but tends to run slower), and the number of epochs (one complete presentation of the entire training data to the network during the training process is called an epoch, an iteration, or one training cycle (Hakin, 1998)). We explored configuration settings by varying combinations of batch size (16 vs. 32), learning rate (5e-5, 3e-5, and 2e-5), and number of epochs (10 vs. 20) and used the validation set to tune the optimal hyperparameter settings that yielded the best overall validation accuracy, which produced the final settings of a batch size of 16, a learning rate of 3e-5, and 10 epochs.

The trained model was then evaluated by the blinded test set (e.g., 1,214 blinded wishes) that the trained BERT model had never encountered before. BERT was analyzed using the Python programming language version 3.8.10 to call the transformers library version 4.11.3 and tensorflow 2.6.0 (details on software platform are available upon request).

Performance of BERT by Predictive Accuracy. Accuracy was evaluated using both *improper scoring* (the percentage of cases correctly classified) and *proper scoring* (the average point-biserial correlation [r_{pb}] between a given human-rated theme' binary value and the BERT-generated probability of a text entry fitting that theme). In the latter case, the average r_{pb} was obtained via Fisher's Z_r statistics (Harrell, 2010; "Scoring rule," 2021). For each prompt, there was a subset of themes with nonzero probabilities generated by BERT and that thus could be tested for their point-biserial correlations with the corresponding binary theme variable as rated by the human coders. Compared to Correct Classification Rate, this correlation constitutes a more finely grained method of evaluating BERT's performance. Even when correct classifications were not made, it would be evidence in BERT's favor if there were a systematic tendency for the probability of being rated with a given theme to be higher in the presence of that theme.

Results

Participant characteristics

The sample included 1253 participants: 285 patients, 349 of their siblings, and 619 in the comparison group (mean age 17, 18, and 19, respectively). The patients were all male, while males made up 48% and 47% of the other 2 groups. Participants resided in a broad cross-section of the United States. One percent of patients, 5% of the siblings and 23% of the comparison group were married. Percentages of Hispanics or Latinos were 9%, 8%, and 20%; percentages of Blacks, 8%, 6%, and 20%. Among patients, 5% were employed, and the rest were unemployed or disabled; in contrast, 42% of the siblings and 61% of the comparison group were employed. Educational levels were varied, with the comparison group having the highest fraction (37%) educated at the bachelor's level or higher. Only 1% of patients or of siblings, but 19% of the comparison group, reported that they or a family member had contracted COVID-19. Comparatively large numbers of participants in all groups reported having help completing the survey: by group, 49%, 26%, and 19%, respectively. Further information is available in the primary publications from this study (Schwartz et al., submitted for publication a, submitted for publication b)

Qualitative Coding Reliability

As reported in the primary papers from this project (Schwartz et al., under review a, b), the mean kappa was 0.77 (*SD*=0.17, range 0.51 to 0.98), reflecting a good level of agreement(Altman, 1999; Landis & Koch, 1977). The best estimated pseudo-R^2 for rater was 0.042 (*p*=0.24), suggesting that the rater effect in coded themes was negligible. Descriptive statistics on proportion of participants whose open-text data reflected various themes are provided in the primary papers from this project.

Examples of Participants’ Wishes

Table 1
Illustrative Examples of Free-Text Entries

Role	Wishes	Definitions	Goals
Patient	Always have a dog, No disease in world, peace	Living life without pain	Going to every baseball stadium. And making a lot of friends
Sibling	I want to be an English teacher. I want to live in a big city. I want to find a partner who loves me very much	Money was plentiful	Doing a degree
Comparison	One wish would be go to Taylor swift next tour. The second, be a millionaire. And the third, have all of taylor swift’s merch.	Quality of life to me means having lived your life in a way that you are proud and know that everything in it was worth it even though it did not seem like it. Also allow yourself to make mistakes and learn and always come back stronger than a 90's trend!	The main things i want to accomplish is get a bachelor's degree in science. Go to the next taylor swift tour, and finally travel the world and see as many of my favorite artists live.

Table 1 provides illustrative examples of free-text entries on what participants wished for, from the combined total of 1214 unique wishes, 480 goals, and 243 definitions randomized into the validation set.

Human Coding Results

Table 2 provides information about the prevalence across the whole sample of Wishes, Goals, and QOL Definition themes. Figure 1 shows the five most prevalent themes by prompt. Financial concerns were prominent across all three

Figure 1. Top five human-coded themes by prompt. The three prompts generated relatively distinct sets of themes, although financial concerns were prominent across all three prompts, and health across two of the three.

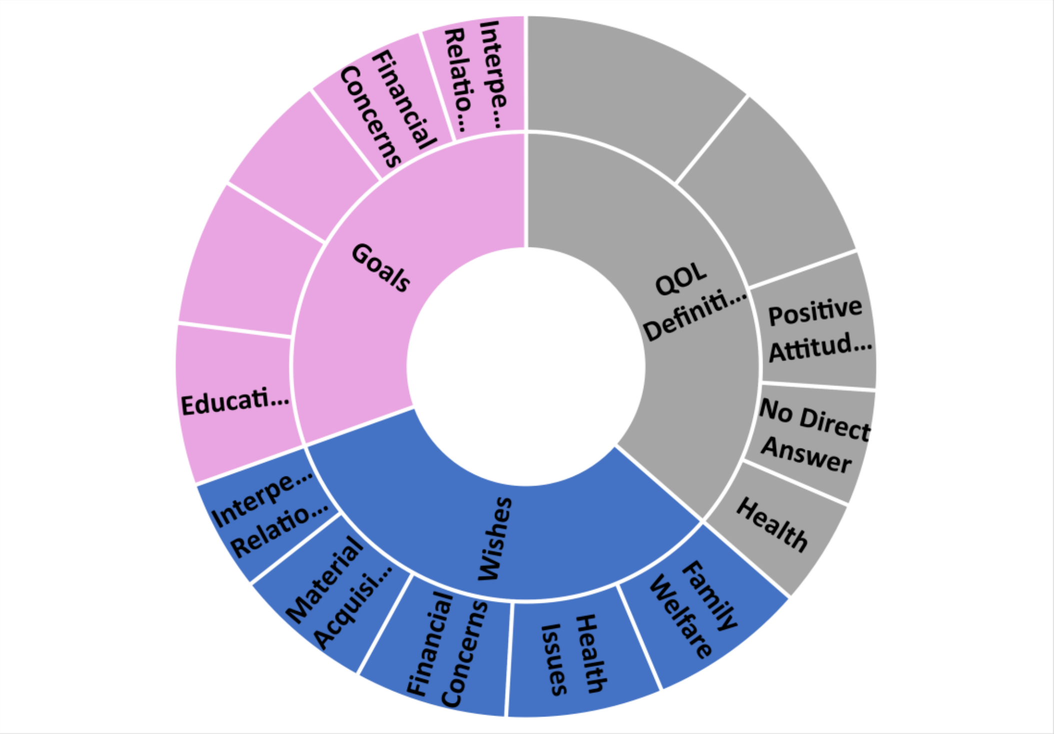


Table 2

Descriptive statistics for human coded themes from open-text prompts, listed from most to least prevalent

Theme	Proportion of sample
Wishes - Family Welfare	.29
Wishes - Health Issues	.29
Wishes - Financial Concerns	.28
Wishes - Material Acquisitions	.26
Wishes - Interpersonal Relationships	.21
Wishes - Travel	.18
Wishes - Work and Unemployment	.17
Wishes - Achievement	.16
Wishes - DMD-Related Goals	.16
Wishes - Societal and Altruistic Concerns	.16
Wishes - Fantasy	.13
Wishes - Leisure Activities	.13
Wishes - Self-Image and Personality	.09
Wishes - Mental Health and Mood State	.09
Wishes - Education	.09
Wishes - COVID-Specific	.08
Wishes - Living Situation, Housing, Neighborhood	.06
Wishes - Health Welfare (Societal)	.06
Wishes - Independent Functioning	.05
Wishes - No Direct Answer	.05
Wishes - Existential Concerns	.03
Wishes - Political Welfare	.02
Wishes - Religious and Spiritual Concerns	.02
Wishes - Accomplishing Chores and Tasks	.02
Wishes - Financial Welfare (Societal)	.02
Wishes - Provider- and Treatment- Related Concerns	.01
Wishes - Problem Resolution	.01
Wishes - Racism	.01
Wishes - Prevention	.01
Wishes - Environmental Welfare	.01
Wishes - Living Situation (Societal)	.01
Wishes - Community Involvement and Voluntarism	.01
Wishes - Disengagement	.00
Wishes - Maintenance	.00
Wishes - Legal and Crime / Safety Concerns	.00
Wishes - Legal and Crime (Societal)	.00
Wishes - Acceptance	.00
Wishes - Drug and Alcohol Use	.00
Wishes - Immigration and Citizenship	.00

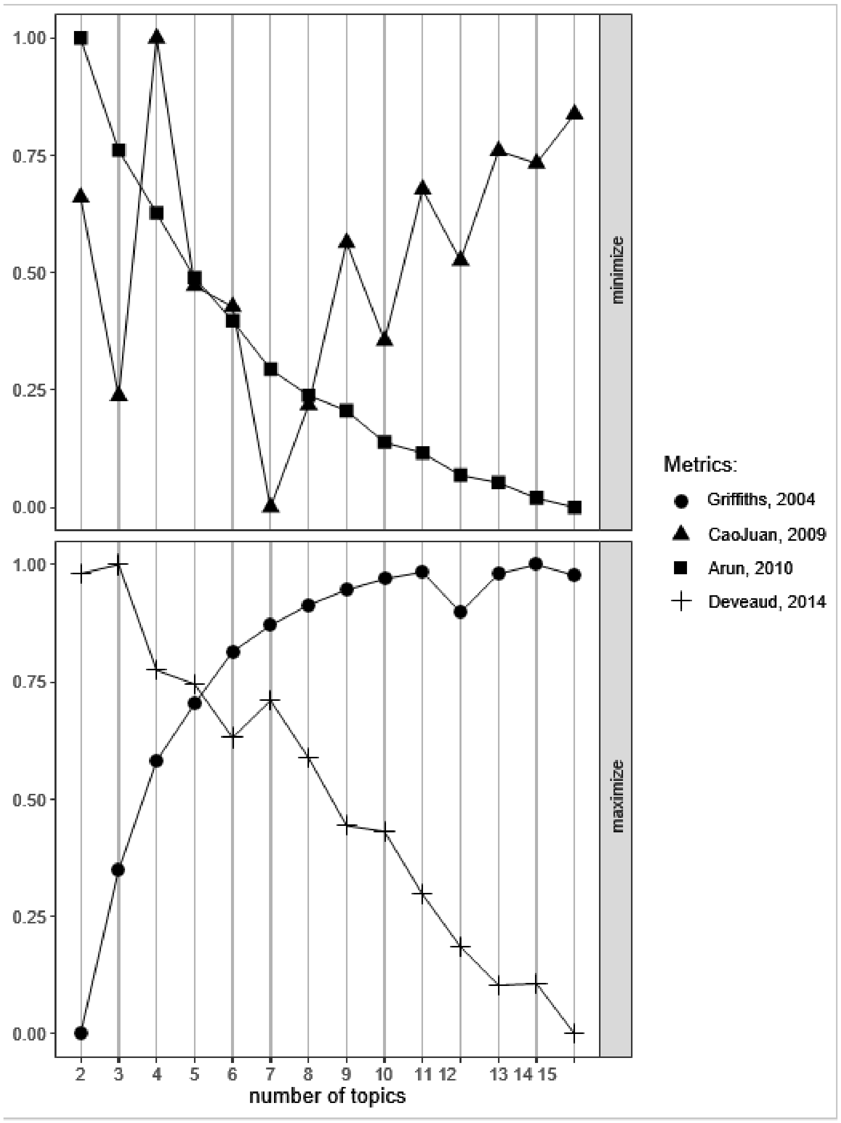
Theme	Proportion of sample
Wishes - Involvement in Community Outreach	.00
Goals - Education	.30
Goals - Work and Unemployment	.27
Goals - Achievement	.23
Goals - Financial Concerns	.23
Goals - Interpersonal Relationships	.20
Goals - No Direct Answer	.19
Goals - Family Welfare	.11
Goals - Mental Health and Mood State	.09
Goals - Living Situation, Housing, Neighborhood	.09
Goals - Health Issues	.07
Goals - Independent Functioning	.07
Goals - Material Acquisitions	.06
Goals - Self-Image and Personality	.06
Goals - Provider- and Treatment- Related Concerns	.04
Goals - Travel	.04
Goals - Societal and Altruistic Concerns	.03
Goals - Community Involvement and Voluntarism	.02
Goals - Accomplishing Chores and Tasks	.02
Goals - Leisure Activities	.02
Goals - Religious and Spiritual Concerns	.02
Goals - Existential Concerns	.02
Goals - DMD-Related Goals	.02
Goals - Acceptance	.01
Goals - Environmental Welfare	.01
Goals - Fantasy	.01
Goals - Health Welfare (Societal)	.01
Goals - Maintenance	.01
Goals - COVID-Specific	.00
Goals - Drug and Alcohol Use	.00
Goals - Prevention	.00
Goals - Disengagement	.00
Goals - Financial Welfare (Societal)	.00
Goals - Immigration and Citizenship	.00
Goals - Involvement in Community Outreach	.00
Goals - Legal and Crime (Societal)	.00
Goals - Legal and Crime / Safety Concerns	.00
Goals - Living Situation (Societal)	.00
Goals - Political Welfare	.00
Goals - Problem Resolution	.00
Goals - Racism	.00

Theme	Proportion of sample
Definition - Circumstances	.44
Definition - Contentment	.35
Definition - Positive Attitude (mental health)	.26
Definition - No Direct Answer	.22
Definition - Health	.20
Definition - Independence	.10
Definition - Personal Growth	.09
Definition - Family / friends	.08
Definition - Contribution	.02
Definition - Treatment-related	.02
Definition - Balance	.01
Definition - Survival	.01
Definition - Problems	.01
Definition - Provider-related	.01
Definition - Reminiscence	.00

LDA Results

How many topics in LDA?. Figure 2 plots the four model-selection metrics by the number of extracted topics. These four metrics provided limited guidance because of their inconsistency. Among the two metrics for which lower score indicates best fit, the Cao et al. (2009) metric suggested either 2 or 9 topics, and the Arun et al. (2010) suggested 10 to15. Among the metrics for which higher score indicates best fit, the Griffiths and Steyvers (2004) metric indicated a model with approximately 8 topics, and the Deveaud et al. (2014) clearly indicated two. We opted to retain 8 as a compromise between the 2 extremes.

Figure 2. Model-selection metrics by the number of extracted topics. Model selection metrics were used to estimate the desired number of topics, using the combined 1,253 statements of validation-set wishes. The top panel plots two metrics that theoretically should behave like a scree plot in an exploratory factor analysis, where the location of the elbow indicates the desired number of topics. Among the two metrics for which lower score indicates best fit, the Cao et al. metric suggested either 2 or 9 topics, and the Arun et al. suggested 10 to15. Among the metrics for which higher score indicates best fit, the Griffiths and Steyvers metric indicated a model with approximately 8 topics, and the Deveaud et al. clearly indicated two.



Topics. Table 3 summarizes the 3 words most strongly associated with each latent topic derived from LDA analysis of respondents' wishes, goals, and QOL definitions. These words inform the interpretation of those topics.

Table 3
Top 3 Words per Latent Topic Derived from LDA Analysis of Respondents' Wishes, Goals and QOL Definitions

Topic						
Wishes	Word1	Prevalence	Word2	Prevalence	Word3	Prevalence
1	like	.08	animals	.04	Walk	.04
2	live	.03	Life	.02	long	.02
3	health	.08	family	.04	love	.03
4	money	.04	DMD	.03	lots	.02
5	world	.07	peace	.04	end	.04
6	travel	.04	brother	.03	money	.02
7	new	.07	house	.05	buy	.03
8	money	.06	healthy	.06	happy	.05
Goals	Word1	Prevalence	Word2	Prevalence	Word3	Prevalence
1	job	.08	business	.04	married	.04
2	debt	.09	Pay	.04	live	.03
3	house	.06	Like	.04	school	.04
4	doctor	.04	Arrangement	.03	doctor arrangement	.03
5	financially	.06	things	.05	stable	.05
6	life	.07	good	.05	family	.03
7	money	.07	happy	.05	make	.05
8	work	.09	degree	.05	study	.04
QOL Definition	Word1	Prevalence	Word2	Prevalence	Word3	Prevalence
1	work	.07	balance	.06	know	.05
2	good	.18	quality	.03	good health	.03
3	happy	.12	Day	.06	healthy	.05
4	material	.11	spiritual	.07	satisfaction	.05
5	able	.09	want	.08	rich	.07
6	living	.13	Live	.11	fullest	.03
7	healthy	.12	happiness	.06	body	.06
8	quality	.08	things	.04	family	.04

For example, Wishes topic 1 includes “like”, “animals”, and “walk”. This topic does not lend itself to easy summary. Topic 2 seems to be related chiefly to living a long life; topic 3, to good health, family, and love; topic 4, to wealth and (finding a cure for) DMD; topic 5, to world peace;   and topic 6, to travel, their brother, and money; topic 7, to worldly possessions; and topic 8 to money, health, and happiness. The fact that even these top three words by topic represent at most 8% of the corresponding text entries, and typically only 4%, makes most of these characterizations tenuous.

For respondents’ goals and QOL definition, the topics do lend themselves to more easy summary. For goals, the eight topics may be loosely characterized, respectively, as ‘finishing school and starting life’, ‘resolving financial debt’, ‘good housing and school’, ‘managing healthcare’, ‘financially stable’, ‘family happiness’, ‘career success’, and ‘college and job prospects’. These top three words by topic represent at most 9% of the corresponding text entries, and on average about 5%.

For QOL definitions, the eight topics may be summarized as ‘work-family balance’, ‘having good health’, ‘happiness & health’, ‘material & spiritual satisfaction’, ‘material wealth’, ‘living life to the fullest’, ‘healthy body’, and ‘provision for family’. These top three words by topic represent at most 18% of the corresponding text entries, and on average about 7%.

BERT Results

Improper Scoring: Correct Classification Rate. Table 4 provides example texts for the goals prompt and shows BERT probabilities for assigning the top five human-coded themes. Grey shading indicates that BERT correctly classified the statement as matching the indicated theme.

Table 4
Examples of BERT's Probabilities that a Given Statement Will Match a Given Theme

Example Text, Goals Prompt	Education	Work & Unemployment	Achievement	Financial Concerns	Interpersonal Relationships
Raising my children to be respectful. Spend as much time with family as possible.	.001	.002	.003	.002	.950 [a]
Live well make a lot of money and retire in asia	.005	.609 [a]	.007	.030	.009
Making a difference, leaving a mark, and achieving my goals	.009	.009	.047	.011	.091
Strive to learn new knowledge	.450	.050	.114	.008	.052

[a] Gray shading that BERT correctly classified the statement as matching the indicated theme.

While BERT correctly identified two themes, it missed others that would have been recognizable as related to one or more themes. For example, “achieving my goals” would have been coded as Achievement by humans but only had a 4.7% probability of such by BERT. Similarly, “strive to learn new knowledge” would have been coded as Education by humans but only had a 45% probability of such by BERT.

Table 5 summarizes the overall accuracy in BERT's predictions for text entries in the validation set, i.e., data that the model had never encountered previously. In the blinded validation set the theme identified by BERT was also identified by humans for 70% of Wishes, 68% of Goals, and 61% of QOL Definition entries, with an overall correct classification rate of 67%. This is despite the fact that BERT could be described as having in most cases “more than one chance." That is, the average statement was rated by human coders as fitting 2.9 themes for Wishes, 2.2 for Goals, and 1.6 for Definitions. BERT thus typically had multiple ways, an average of 2.6, in which its classification could conceivably match some human-coded theme.

Table 5
DistillBERT's Predictive Accuracy

		Training Set		Blinded Validation Set	
Prompt	*n* codes	*n* entries	Accuracy	*n* entries	Accuracy
Wishes	40	399	100%	1214	70%
Goals	40	160	100%	480	68%
QOL Definition	15	139	100%	545	61%

Proper Scoring: Human-BERT Correlation. Table 6 shows the average correlations among themes coded by humans and BERT, separately by prompt. For themes within all three prompts, the algorithm's probabilities generally correlated only moderately with the binary theme variable, with average r_{pb} per prompt in the 0.3-0.4 range and an overall r_{pb} of 0.34 (Table 6). These correlations reflect a relatively low overall explained variance of 0.12, with more variance explained for Goals (R^2=0.14) than for Wishes (R^2=0.11) or QOL Definition (R^2=0.10).

Table 6
Correlations Among Themes Coded by Humans vs. BERT

Prompt	*n* Comparisons	Mean r_{pb}*	Minimum r_{pb}*	Maximum r_{pb}*
Wishes (*n* =1,207)	30	0.33	-0.01	0.85
Goals (*n* = 478)	21	0.37	-0.01	0.70
Definitions (*n* = 243)	11	0.32	0.06	0.57
Total (weighted by *n* Comparisons)	62	0.34	-0.01	0.85

Note. r_{pb} = point-biserial correlation coefficient

Relative Efficiency. Considering all of the time needed for training and scoring the open-text data, the three methods took very similar amounts of time. LDA and BERT took approximately 1.15 minutes per training sample (on a 64-bit workstation with a 6-core Intel Xeon CPU at 2.40 GHz and 32 GiB of memory running Ubuntu Linux version 20.04, no GPU was utilized). By comparison, human raters can code one entry at an average rate of 1.18 minutes. After removing time for training and programming, LDA took about 8 seconds per entry, and BERT took 4. After removing time for training, the human raters took about 52 seconds per entry.

Discussion

The present study is, to our knowledge, the first to compare human coding to two NLP methods - LDA and BERT - for analyzing large-scale qualitative data. Table 7 summarizes the features of the three methods. Compared to human coders, LDA in this study did not yield easily interpretable themes. LDA output is difficult to summarize in meaningful ways, partially because the same word, phrase, or theme can appear multiple times across latent topics. BERT has the potential to be more useful because it can be trained to recognize topics or themes already deemed meaningful by humans. Nonetheless, BERT accurately identified only about two thirds of statements that it had never encountered previously in training, despite having on average 2.6 themes that humans had coded for any given text entry. Moreover, the more sensitive point-biserial correlation showed an average explained variance of 12% per theme. Because LDA and BERT require specialized knowledge and software, their feasibility and accessibility may be limited for researchers without such access.

Table 7
Summary of Text-Analysis Methods

		Method	
Feature	Humans	LDA	BERT
Yields interpretable themes	√		
Training required	√		√
High hourly cost		√	√
Specialized knowledge required	(√)	√	√
Special Software required	(√)	√	√
Scalable to big data (*n*>100K)		√	√

Our findings on LDA are different from Baumer et al.'s (2017) results and from the impressive results found in the wider literature on LDA, in which LDA is able to extract coherent and meaningful themes. The seminal paper on LDA (Blei, Ng, & Jordan, 2003) showed that LDA extracted meaningful and unique topics from over 16 thousand newswire articles. LDA also successfully found themes from over 40 thousand entries of chapter-length reading materials for students (Steyvers & Griffiths, 2007) or scientific abstracts (Griffiths & Steyvers, 2004). Like any statistical procedure, LDA's performance depends on the contents in the input data. Our findings suggest that LDA does not perform well in the context of relatively brief open-text entries.

Other researchers may get different results if BERT is applied after a much larger training set (i.e., longer open-text entries, far fewer themes, and many more entries per theme). For example, Murarka, Radhakrishnan,

& Ravichandran (2020) analyzed 17,000 social-media posts and achieved 80% accuracy in classifying posts into one of five specific mental-health outcomes. In contrast, our data derived from three relatively broad prompts about wishes, QOL definition, and goals, and were human-coded into 95 themes. This is a more complex task that may draw on empathy and life experience. One other limitation of BERT in text classification is that it requires training of human coders to generate coded data that can be used to train BERT. Thus, the highest cost of human coders (i.e., the training and adjudication period) would need to be included in the overall cost of BERT.

It is worth noting that any successful implementation of BERT could be reused once trained. In our case, for example, if it had a greater accuracy (e.g., > 80% similar to (Murarka et al., 2020)), our BERT model could have been applied to classify the wishes and aspirations of people who post online about Muscular Dystrophy. Also, because our data include a comparison group, the neural-network weights devised might be applied to understand the aspirations of individuals from the general population. This reusability may offset the initial cost of training BERT.

Humans provided reliable, valid, and cost-effective coding in the web-based context with relatively short text entries. On average, they took only two seconds longer than LDA or BERT per open-text entry. Of note, the present study included coding of approximately 3,000 open-text entries. Thus, scaling up to larger data sets and longer text entries might be feasible for motivated and compensated human coders. We have not evaluated the three methods in processing other qualitative data such as interview transcripts. Future research might compare the three methods in the context of hour-long interview transcripts, where BERT's advantages may be more apparent.

This study has many advantages, including a robust sample with good quality data on multiple prompts. Nonetheless, the limitations of the study must be acknowledged. First, there is considerable uncertainty in the LDA results, as seen in the unexpected patterns in two of the four model-selection metrics. Also, the current BERT model only predicts one code at a time, even though it is capable of predicting multiple categories. This was a crude but reasonable and practical beginning of this line of inquiry. Future research should examine LDA's results and interpretability to provide guidance as to when the method is most appropriate. Future BERT modeling can go onto multi-class task. Another limitation of the present work is that the computed correlations between BERT and humans are likely attenuated by the continuous-binary pairing. BERT's average r_{pb} of 0.34 would thus likely be somewhat larger, and would translate to more than our documented 12% explained variance. However, even if that 12% were tripled, it would not seem enough to justify replacing human raters with this algorithm.

Conclusions

In summary, LDA and BERT provide potentially viable approaches to analyzing large-scale qualitative data, but both have limitations. When text entries are short, LDA yields latent topics that are hard to interpret. BERT accurately identified only about two thirds of new statements even given multiple opportunities. Moreover, the probabilities it assigned showed unsatisfactory correlations with the binary theme variables in question. Humans provided reliable and cost-effective coding in the web-based context. Future research should examine NLP's predictive accuracy given different contexts and quantities of training data.

Author Note. Corresponding author: Carolyn Schwartz, Sc.D., DeltaQuest Foundation, Inc., Concord, MA, email: carolyn.schwartz@deltaquest.org.
Acknowledgments. We are grateful to the participants themselves who provided data for this project; to Alexander Denby, Julia B. Lewis, Emily Kane, Matyas Csiki-Fejer, for their careful efforts coding the open-text data; and to Sarepta Therapeutics for support of data collection for this work.

References

Agaronnik, N., Lindvall, C., El-Jawahri, A., He, W., & Iezzoni, L. (2020). Use of natural language processing to assess frequency of functional status documentation for patients newly diagnosed with colorectal cancer. *JAMA Oncology, 6*, 1628-1630.

Altman, D. G. (1999). *Practical statistics for medical research*. New York: Chapman & Hall/CRC Press.

Arun, R., Suresh, V., Veni Madhavan, C. E., & Murthy, N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In M. J. Zaki, J. X. Yu, B. Ravindran, & V. Pudi (Eds.), *Advances in Knowledge Discovery and Data Mining*. (pp. 391 - 402). Heidelberg: Springer Berlin.

Atkinson, T. M. (2018). *Latent dirichlet allocation in discovering goals in patients undergoing bladder cancer surgery*. Paper presented at the Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics.

Barry, M. J., & Edgman-Levitan, S. (2012). Shared decision making—The pinnacle patient-centered care. *New England Journal of Medicine, 366*, 780-781. doi:10.1056/NEJMp1109283

Baumer, E. P., Mimno, D., Guha, S., Quan, E., & Gay, G. K. (2017). Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology, 68*, 1397-1410.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993-1022.

Can, D., Marín, R., Georgiou, P., Imel, Z., Atkins, D., & Narayanan, S. (2016). "It sounds like...": A natural language processing approach to detecting counselor reflections in motivational interviewing. *Journal of Counseling Psychology, 63*, 343-350.

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing - 16th European Symposium on Artificial Neural Networks, 72*, 1775 - 1781. doi:10.1016/j.neucom.2008.06.011

Cappelleri, J. C., Zou, K. H., Bushmakin, A. G., Alvir, J. M. J., Alemayehu, D., & Symonds, T. (2013). Development of a patient-reported outcome. In *Patient-reported Outcomes: Measurement, Implementation and Interpretation* (pp. 21-29). Boca Raton, FL: CRC Press.

Chen, P.-H., Zafar, H., Galperin-Aizenberg, M., & Cook, T. (2018). Integrating natural language processing and machine learning algorithms to categorize oncologic response in radiology reports. *Journal of Digital Imaging, 31*, 178-184. doi:10.1007/s10278-017-0027-x

Ciafaloni, E., Fox, D. J., Pandya, S., Westfield, C. P., Puzhankara, S., Romitti, P. A., . . . Miller, L. A. (2009). Delayed diagnosis in Duchenne Muscular Dystrophy: Data from the Muscular Dystrophy Surveillance, Tracking, and Research Network (MD STARnet). *The Journal of Pediatrics, 155*, 380-385.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Deveaud, R., SanJuan, É., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique, 17*, 61-84. doi:10.3166/dn.17.1.61-84

Devlin, J., & Chang, M.-W. (2018). Open Sourcing BERT: State-of-the-Art Pre-training for Natural language processing. *Google AI Blog, 2*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dönges, J. (2009). What Your Choice of Words Says about Your Personality. *Scientific American Mind 20*, 4, 14-15.

Ferrans, C. E. (2005). Definitions and conceptual models of quality of life. In J. Lipscomb, C. D. Gotay, & C. Snyder (Eds.), *Outcomes Assessment in Cancer: Measures, Methods, and Applications* (pp. 14-30). Cambridge, UK: Cambridge University Press.

Finch, W. H., Hernández Finch, M., McIntosh, C., & Braun, C. (2018). The use of topic modeling with latent Dirichlet analysis with open-ended survey items. *Translational Issues in Psychological Science, 4*, 403-424.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*, 378-382.

Gonzalez-Hernandez, G., Sarker, A., O'Connor, K., & Savova, G. (2017). Capturing the patient's perspective: A review of advances in natural language processing of health-related text. *Yearbook of Medical Informatics, 26*, 214-227. doi:10.15265/IY-2017-029

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(Suppl 1), 5228-5235.

Guetterman, T. C., Chang, T., DeJonckheere, M., Basu, T., Scruggs, E., & Vydiswaran, V. V. (2018). Augmenting qualitative text analysis with natural language processing: Methodological study. *Journal of Medical Internet Research, 20*, e9702.

Hakin, S., (1998). *Neural networks: A comprehensive foundation* (Second ed.). Hoboken, NJ: Prentice Hall PTR.

Harrell, F. E. (2010). *Regression modeling strategies*. New York: Springer.

Hugging Face. (2021). The AI community building the future. *URL* https://huggingface.co

Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., & Pennebaker, J. W. (2011). Language Style Matching Predicts Relationship Initiation and Stability. *Psychological Science, 22*, 39-44. doi:10.1177/0956797610392928

Kebede, S. (2016). Ask patients "What matters to you?" rather than "What's the matter?". *BMJ, 354*.

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. arXiv:1412.6980.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Leeson, W., Resnick, A., Alexander, D., & Rovers, J. (2019). Natural language processing (NLP) in qualitative public health research: a proof of concept study. *International Journal of Qualitative Methods, 18,* 1-9.

Li, Y., Rapkin, B., Atkinson, T. M., Schofield, E., & Bochner, B. H. (2019). Leveraging latent dirichlet allocation in processing free-text personal goals among patients undergoing bladder cancer surgery. *Quality of Life Research, 28*, 1441-1455. doi:10.1007/s11136-019-02132-w

Li, Y., & Rapkin, B. D. (2009). Classification and regression tree analysis to identify complex cognitive paths underlying quality of life response shifts: A study of individuals living with HIV/AIDS. *Journal of Clinical Epidemiology, 62*, 1138-1147.

Hugging Face (2021). List of alternative algorithms derived from the full BERT technology.

URL https://huggingface.co/transformers/pretrained_models.html

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Moyers, T., Martin, T., Catley, D., Harris, K. J., & Ahluwalia, J. S. (2003). Assessing the integrity of motivational interviewing interventions: Reliability of the motivational interviewing skills code. *Behavioural and Cognitive Psychotherapy, 31*, 177.

Murarka, A., Radhakrishnan, B., & Ravichandran, S. (2020). Detection and classification of mental illnesses on social media using RoBERTa. *arXiv preprint arXiv:2011.11226*.

Nereo, N. E., & Hinton, V. J. (2003). Three wishes and psychological functioning in boys with Duchenne muscular dystrophy. *Journal of developmental and behavioral pediatrics: JDBP, 24*, 96.

Nikita, M. (2016). ldatuning: Tuning of the latent dirichlet allocation models parameters. In: R package version 0.2.0.

Pane, M., Lombardo, M. E., Alfieri, P., D'Amico, A., Bianco, F., Vasco, G., . . . Ricotti, V. (2012). Attention deficit hyperactivity disorder and cognitive function in Duchenne muscular dystrophy: Phenotype-genotype correlation. *The Journal of Pediatrics, 161,* 705-709. e701.

Parker, S. T. (2020). Estimating nonfatal gunshot injury locations with natural language processing and machine learning models. *JAMA Network Open, 3*, e2020664-e2020664.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825 - 2830.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates, 71*.

Perrin, A. (2001). The CodeRead System: Using natural language processing to automate coding of qualitative data. *Social Science Computer Review, 19*, 213-220.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., . . . Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*, 1-67.

Rapkin, B. D., & Schwartz, C. E. (2004). Toward a theoretical model of quality-of-life appraisal: Implications of findings from studies of response shift. *Health and Quality of Life Outcomes, 2*, 14.

Receptiviti. (2021). LIWC. *URL* https://www.receptiviti.com/liwc

Reyes, S. (2019). Multi-class text classification (TFIDF). *URL* https://www.kaggle.com/selener/multi-class-text-classification-tfidf

Schwartz, C. E., Biletch, E., Stuart, R. B. B., & Rapkin, B. D. Patient aspirations in the context of Duchenne Muscular Dystrophy: A mixed-methods case-control study. *Under review*.

Schwartz, C. E., Biletch, E., Stuart, R. B. B., & Rapkin, B. D. Sibling aspirations in the context of Duchenne Muscular Dystrophy: A mixed-methods case-control study. *Under review*.

Schwartz, C. E., & Revicki, D. A. (2012). Mixing methods and blending paradigms: Some considerations for future research. *Quality of Life Research, 21*, 375-376. doi:10.1007/s11136-012-0124-8

Scoring rule. (2021). In Wikipedia. https://en.wikipedia.org/wiki/Scoring_rule#Proper_scoring_rules

Skaljic, M., Patel, I. H., Pellegrini, A. M., Castro, V. M., Perlis, R. H., & Gordon, D. D. (2019). Prevalence of financial considerations documented in primary care encounters as identified by natural language processing methods. *JAMA Network Open, 2*, e1910399-e1910399.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of Latent Semantic Analysis*. New York: Psychology Press, Taylor & Francis Group.

Tanana, M., Hallgren, K. A., Imel, Z. E., Atkins, D. C., & Srikumar, V. (2016). A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of Substance Abuse Treatment, 65*, 43-50.

Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. *arXiv preprint arXiv:1905.05950*.

Trotter II, R. T. (2012). Qualitative research sample design and sample size: Resolving and unresolved issues and inferential imperatives. *Preventive Medicine, 55,* 398-400.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998-6008.

SN Computer Science (2021) 2:160
https://doi.org/10.1007/s42979-021-00592-x

REVIEW ARTICLE

Machine Learning: Algorithms, Real-World Applications and Research Directions

Iqbal H. Sarker[1,2]

Received: 27 January 2021 / Accepted: 12 March 2021 / Published online: 22 March 2021

Abstract
In the current age of the Fourth Industrial Revolution (4*IR* or Industry 4.0), the digital world has a wealth of data, such as Internet of Things (IoT) data, cybersecurity data, mobile data, business data, social media data, health data, etc. To intelligently analyze these data and develop the corresponding *smart and automated* applications, the knowledge of artificial intelligence (AI), particularly, *machine learning (ML)* is the key. Various types of machine learning algorithms such as supervised, unsupervised, semi-supervised, and reinforcement learning exist in the area. Besides, the *deep learning*, which is part of a broader family of machine learning methods, can intelligently analyze the data on a large scale. In this paper, we present a comprehensive view on these *machine learning algorithms* that can be applied to enhance the intelligence and the capabilities of an application. Thus, this study's key contribution is explaining the principles of different machine learning techniques and their applicability in various real-world *application* domains, such as cybersecurity systems, smart cities, healthcare, e-commerce, agriculture, and many more. We also highlight the challenges and potential *research directions* based on our study. Overall, this paper aims to serve as a reference point for both academia and industry professionals as well as for decision-makers in various real-world situations and application areas, particularly from the technical point of view.

Keywords Machine learning · Deep learning · Artificial intelligence · Data science · Data-driven decision-making · Predictive analytics · Intelligent applications

Introduction

We live in the age of data, where everything around us is connected to a data source, and everything in our lives is digitally recorded [21, 103]. For instance, the current electronic world has a wealth of various kinds of data, such as the Internet of Things (IoT) data, cybersecurity data, smart city data, business data, smartphone data, social media data, health data, COVID-19 data, and many more. The data can be structured, semi-structured, or unstructured, discussed briefly in Sect. "Types of Real-World Data and Machine Learning Techniques", which is increasing day-by-day. Extracting insights from these data can be used to build various intelligent applications in the relevant domains. For instance, to build a data-driven automated and intelligent cybersecurity system, the relevant cybersecurity data can be used [105]; to build personalized context-aware smart mobile applications, the relevant mobile data can be used [103], and so on. Thus, the data management tools and techniques having the capability of *extracting insights or useful knowledge* from the data in a timely and intelligent way is urgently needed, on which the real-world applications are based.

Artificial intelligence (AI), particularly, *machine learning (ML)* have grown rapidly in recent years in the context of data analysis and computing that typically allows the applications to function in an intelligent manner [95]. ML usually provides systems with the ability to learn and enhance from experience automatically without being specifically programmed and is generally referred to as the most popular

This article is part of the topical collection "Advances in Computational Approaches for Artificial Intelligence, Image Processing, IoT and Cloud Applications" guest edited by Bhanu Prakash K N and M. Shivakumar.

✉ Iqbal H. Sarker
msarker@swin.edu.au

[1] Swinburne University of Technology, Melbourne, VIC 3122, Australia

[2] Department of Computer Science and Engineering, Chittagong University of Engineering & Technology, 4349 Chattogram, Bangladesh

latest technologies in the fourth industrial revolution (4*IR* or Industry 4.0) [103, 105]. “Industry 4.0” [114] is typically the ongoing automation of conventional manufacturing and industrial practices, including exploratory data processing, using new smart technologies such as machine learning automation. Thus, to intelligently analyze these data and to develop the corresponding real-world applications, *machine learning algorithms* is the key. The learning algorithms can be categorized into four major types, such as supervised, unsupervised, semi-supervised, and reinforcement learning in the area [75], discussed briefly in Sect. “Types of Real-World Data and MachineLearning Techniques”. The popularity of these approaches to learning is increasing day-by-day, which is shown in Fig. 1, based on data collected from Google Trends [4] over the last five years. The *x-axis* of the figure indicates the specific dates and the corresponding popularity score within the range of 0 (*minimum*) to 100 (*maximum*) has been shown in *y-axis*. According to Fig. 1, the popularity indication values for these learning types are low in 2015 and are increasing day by day. These statistics motivate us to study on *machine learning* in this paper, which can play an important role in the real-world through Industry 4.0 automation.

In general, the effectiveness and the efficiency of a *machine learning* solution depend on the nature and characteristics of *data* and the performance of the *learning algorithms*. In the area of machine learning algorithms, classification analysis, regression, data clustering, feature engineering and dimensionality reduction, association rule learning, or reinforcement learning techniques exist to effectively build data-driven systems [41, 125]. Besides, *deep learning* originated from the artificial neural network that can be used to intelligently analyze data, which is known as part of a wider family of machine learning approaches [96]. Thus, selecting a proper learning algorithm that is suitable for the target application in a particular domain is challenging. The reason is that the purpose of different learning algorithms is different, even the outcome of different learning algorithms in a similar category may vary depending on the data characteristics [106]. Thus, it is important to understand the principles of various machine learning algorithms and their applicability to apply in various real-world application areas, such as IoT systems, cybersecurity services, business and recommendation systems, smart cities, healthcare and COVID-19, context-aware systems, sustainable agriculture, and many more that are explained briefly in Sect. “Applications of Machine Learning”.

Based on the importance and potentiality of “Machine Learning” to analyze the data mentioned above, in this paper, we provide a comprehensive view on various types of *machine learning algorithms* that can be applied to enhance the intelligence and the capabilities of an application. Thus, the key contribution of this study is explaining the principles and potentiality of different machine learning techniques, and their applicability in various real-world application areas mentioned earlier. The purpose of this paper is, therefore, to provide a basic guide for those *academia and industry* people who want to study, research, and develop data-driven automated and intelligent systems in the relevant areas based on machine learning techniques.

The key contributions of this paper are listed as follows:

– To define the scope of our study by taking into account the nature and characteristics of various types of real-world data and the capabilities of various learning techniques.
– To provide a comprehensive view on machine learning algorithms that can be applied to enhance the intelligence and capabilities of a data-driven application.

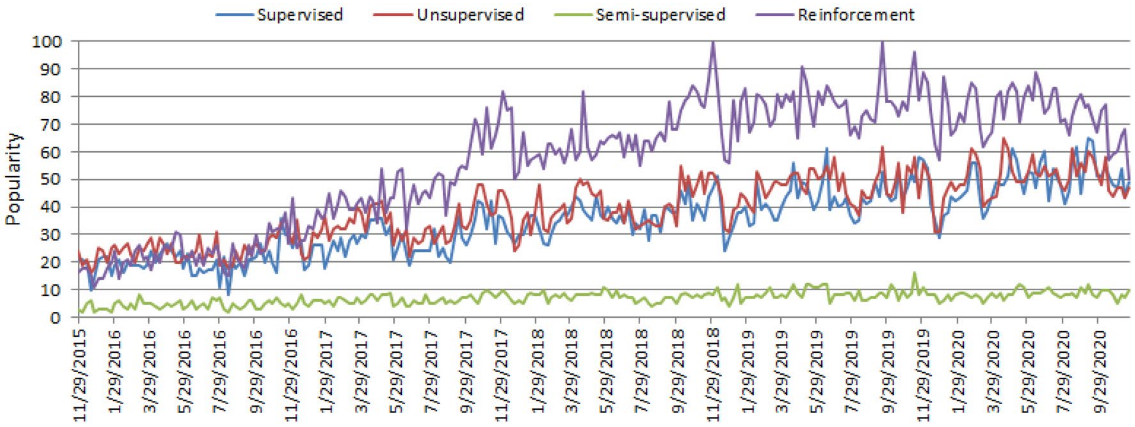


Fig. 1 The worldwide popularity score of various types of ML algorithms (supervised, unsupervised, semi-supervised, and reinforcement) in a range of 0 (min) to 100 (max) over time where x-axis represents the timestamp information and y-axis represents the corresponding score

- To discuss the applicability of machine learning-based solutions in various real-world application domains.
- To highlight and summarize the potential research directions within the scope of our study for intelligent data analysis and services.

The rest of the paper is organized as follows. The next section presents the types of data and machine learning algorithms in a broader sense and defines the scope of our study. We briefly discuss and explain different machine learning algorithms in the subsequent section followed by which various real-world application areas based on machine learning algorithms are discussed and summarized. In the penultimate section, we highlight several research issues and potential future directions, and the final section concludes this paper.

Types of Real-World Data and Machine Learning Techniques

Machine learning algorithms typically consume and process data to learn the related patterns about individuals, business processes, transactions, events, and so on. In the following, we discuss various types of real-world data as well as categories of machine learning algorithms.

Types of Real-World Data

Usually, the availability of data is considered as the key to construct a machine learning model or data-driven real-world systems [103, 105]. Data can be of various forms, such as structured, semi-structured, or unstructured [41, 72]. Besides, the "metadata" is another type that typically represents data about the data. In the following, we briefly discuss these types of data.

- *Structured:* It has a well-defined structure, conforms to a data model following a standard order, which is highly organized and easily accessed, and used by an entity or a computer program. In well-defined schemes, such as relational databases, structured data are typically stored, i.e., in a tabular format. For instance, names, dates, addresses, credit card numbers, stock information, geolocation, etc. are examples of structured data.
- *Unstructured:* On the other hand, there is no pre-defined format or organization for unstructured data, making it much more difficult to capture, process, and analyze, mostly containing text and multimedia material. For example, sensor data, emails, blog entries, wikis, and word processing documents, PDF files, audio files, videos, images, presentations, web pages, and many other types of business documents can be considered as unstructured data.
- *Semi-structured:* Semi-structured data are not stored in a relational database like the structured data mentioned above, but it does have certain organizational properties that make it easier to analyze. HTML, XML, JSON documents, NoSQL databases, etc., are some examples of semi-structured data.
- *Metadata:* It is not the normal form of data, but "data about data". The primary difference between "data" and "metadata" is that data are simply the material that can classify, measure, or even document something relative to an organization's data properties. On the other hand, metadata describes the relevant data information, giving it more significance for data users. A basic example of a document's metadata might be the author, file size, date generated by the document, keywords to define the document, etc.

In the area of machine learning and data science, researchers use various widely used datasets for different purposes. These are, for example, cybersecurity datasets such as NSL-KDD [119], UNSW-NB15 [76], ISCX'12 [1], CIC-DDoS2019 [2], Bot-IoT [59], etc., smartphone datasets such as phone call logs [84, 101], SMS Log [29], mobile application usages logs [137] [117], mobile phone notification logs [73] etc., IoT data [16, 57, 62], agriculture and e-commerce data [120, 138], health data such as heart disease [92], diabetes mellitus [83, 134], COVID-19 [43, 74], etc., and many more in various application domains. The data can be in different types discussed above, which may vary from application to application in the real world. To analyze such data in a particular problem domain, and to extract the insights or useful knowledge from the data for building the real-world intelligent applications, different types of machine learning techniques can be used according to their learning capabilities, which is discussed in the following.

Types of Machine Learning Techniques

Machine Learning algorithms are mainly divided into four categories: Supervised learning, Unsupervised learning, Semi-supervised learning, and Reinforcement learning [75], as shown in Fig. 2. In the following, we briefly discuss each type of learning technique with the scope of their applicability to solve real-world problems.

- *Supervised:* Supervised learning is typically the task of machine learning to learn a function that maps an input to an output based on sample input-output pairs [41]. It uses labeled training data and a collection of training examples to infer a function. Supervised learning is carried out when certain goals are identified to be accom-

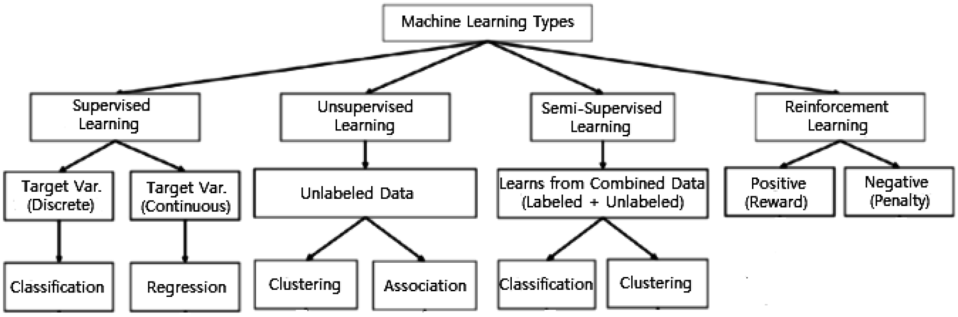


Fig. 2 Various types of machine learning techniques

plished from a certain set of inputs [105], i.e., a *task-driven approach*. The most common supervised tasks are "classification" that separates the data, and "regression" that fits the data. For instance, predicting the class label or sentiment of a piece of text, like a tweet or a product review, i.e., text classification, is an example of supervised learning.

- *Unsupervised:* Unsupervised learning analyzes unlabeled datasets without the need for human interference, i.e., a *data-driven process* [41]. This is widely used for extracting generative features, identifying meaningful trends and structures, groupings in results, and exploratory purposes. The most common unsupervised learning tasks are clustering, density estimation, feature learning, dimensionality reduction, finding association rules, anomaly detection, etc.
- *Semi-supervised:* Semi-supervised learning can be defined as a *hybridization* of the above-mentioned supervised and unsupervised methods, as it operates on both labeled and unlabeled data [41, 105]. Thus, it falls between learning "without supervision" and learning "with supervision". In the real world, labeled data could be rare in several contexts, and unlabeled data are numerous, where semi-supervised learning is useful [75]. The ultimate goal of a semi-supervised learning model is to provide a better outcome for prediction than that produced using the labeled data alone from the model. Some application areas where semi-supervised learning is used include machine translation, fraud detection, labeling data and text classification.
- *Reinforcement:* Reinforcement learning is a type of machine learning algorithm that enables software agents and machines to automatically evaluate the optimal behavior in a particular context or environment to improve its efficiency [52], i.e., an *environment-driven approach*. This type of learning is based on reward or penalty, and its ultimate goal is to use insights obtained from environmental activists to take action to increase the reward or minimize the risk [75]. It is a powerful tool for training AI models that can help increase automation or optimize the operational efficiency of sophisticated systems such as robotics, autonomous driving tasks, manufacturing and supply chain logistics, however, not preferable to use it for solving the basic or straightforward problems.

Thus, to build effective models in various application areas different types of machine learning techniques can play a significant role according to their learning capabilities, depending on the nature of the data discussed earlier, and the target outcome. In Table 1, we summarize various types of machine learning techniques with examples. In the following, we provide a comprehensive view of machine learning algorithms that can be applied to enhance the intelligence and capabilities of a data-driven application.

Table 1 Various types of machine learning techniques with examples

Learning type	Model building	Examples
Supervised	Algorithms or models learn from labeled data (task-driven approach)	Classification, regression
Unsupervised	Algorithms or models learn from unlabeled data (Data-Driven Approach)	Clustering, associations, dimensionality reduction
Semi-supervised	Models are built using combined data (labeled + unlabeled)	Classification, clustering
Reinforcement	Models are based on reward or penalty (environment-driven approach)	Classification, control

Machine Learning Tasks and Algorithms

In this section, we discuss various machine learning algorithms that include classification analysis, regression analysis, data clustering, association rule learning, feature engineering for dimensionality reduction, as well as deep learning methods. A general structure of a machine learning-based predictive model has been shown in Fig. 3, where the model is trained from historical data in phase 1 and the outcome is generated in phase 2 for the new test data.

Classification Analysis

Classification is regarded as a supervised learning method in machine learning, referring to a problem of predictive modeling as well, where a class label is predicted for a given example [41]. Mathematically, it maps a function (*f*) from input variables (*X*) to output variables (*Y*) as target, label or categories. To predict the class of given data points, it can be carried out on structured or unstructured data. For example, spam detection such as "spam" and "not spam" in email service providers can be a classification problem. In the following, we summarize the common classification problems.

- *Binary classification:* It refers to the classification tasks having two class labels such as "true and false" or "yes and no" [41]. In such binary classification tasks, one class could be the normal state, while the abnormal state could be another class. For instance, "cancer not detected" is the normal state of a task that involves a medical test, and "cancer detected" could be considered as the abnormal state. Similarly, "spam" and "not spam" in the above example of email service providers are considered as binary classification.
- *Multiclass classification:* Traditionally, this refers to those classification tasks having more than two class labels [41]. The multiclass classification does not have the principle of normal and abnormal outcomes, unlike binary classification tasks. Instead, within a range of specified classes, examples are classified as belonging to one. For example, it can be a multiclass classification task to classify various types of network attacks in the NSL-KDD [119] dataset, where the attack categories are classified into four class labels, such as DoS (Denial of Service Attack), U2R (User to Root Attack), R2L (Root to Local Attack), and Probing Attack.
- *Multi-label classification:* In machine learning, multi-label classification is an important consideration where an example is associated with several classes or labels. Thus, it is a generalization of multiclass classification, where the classes involved in the problem are hierarchically structured, and each example may simultaneously belong to more than one class in each hierarchical level, e.g., multi-level text classification. For instance, Google news can be presented under the categories of a "city name", "technology", or "latest news", etc. Multi-label classification includes advanced machine learning algorithms that support predicting various mutually non-exclusive classes or labels, unlike traditional classification tasks where class labels are mutually exclusive [82].

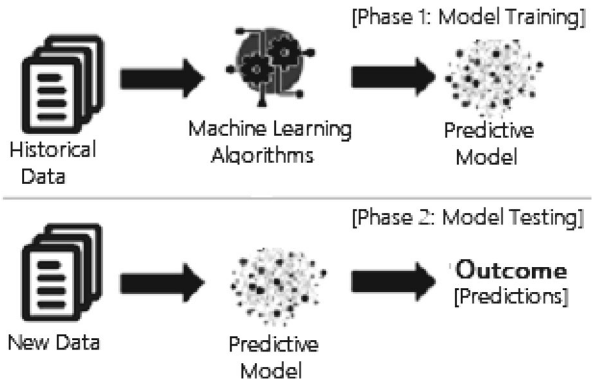


Fig. 3 A general structure of a machine learning based predictive model considering both the training and testing phase

Many classification algorithms have been proposed in the machine learning and data science literature [41, 125]. In the following, we summarize the most common and popular methods that are used widely in various application areas.

- *Naive Bayes (NB):* The naive Bayes algorithm is based on the Bayes' theorem with the assumption of independence between each pair of features [51]. It works well and can be used for both binary and multi-class categories in many real-world situations, such as document or text classification, spam filtering, etc. To effectively classify the noisy instances in the data and to construct a robust prediction model, the NB classifier can be used [94]. The key benefit is that, compared to more sophisticated approaches, it needs a small amount of training data to estimate the necessary parameters and quickly [82]. However, its performance may affect due to its strong assumptions on features independence. Gaussian, Multinomial, Complement, Bernoulli, and Categorical are the common variants of NB classifier [82].
- *Linear Discriminant Analysis (LDA):* Linear Discriminant Analysis (LDA) is a linear decision boundary classifier created by fitting class conditional densities to data and applying Bayes' rule [51, 82]. This method is also known as a generalization of Fisher's linear discriminant, which projects a given dataset into a lower-dimensional space, i.e., a reduction of dimensionality that minimizes

the complexity of the model or reduces the resulting model's computational costs. The standard LDA model usually suits each class with a Gaussian density, assuming that all classes share the same covariance matrix [82]. LDA is closely related to ANOVA (analysis of variance) and regression analysis, which seek to express one dependent variable as a linear combination of other features or measurements.

- *Logistic regression (LR):* Another common probabilistic based statistical model used to solve classification issues in machine learning is Logistic Regression (LR) [64]. Logistic regression typically uses a logistic function to estimate the probabilities, which is also referred to as the mathematically defined sigmoid function in Eq. 1. It can overfit high-dimensional datasets and works well when the dataset can be separated linearly. The regularization (L1 and L2) techniques [82] can be used to avoid over-fitting in such scenarios. The assumption of linearity between the dependent and independent variables is considered as a major drawback of Logistic Regression. It can be used for both classification and regression problems, but it is more commonly used for classification.

$$g(z) = \frac{1}{1 + \exp(-z)}. \tag{1}$$

- *K-nearest neighbors (KNN):* K-Nearest Neighbors (KNN) [9] is an "instance-based learning" or non-generalizing learning, also known as a "lazy learning" algorithm. It does not focus on constructing a general internal model; instead, it stores all instances corresponding to training data in n-dimensional space. KNN uses data and classifies new data points based on similarity measures (e.g., Euclidean distance function) [82]. Classification is computed from a simple majority vote of the k nearest neighbors of each point. It is quite robust to noisy training data, and accuracy depends on the data quality. The biggest issue with KNN is to choose the optimal number of neighbors to be considered. KNN can be used both for classification as well as regression.
- *Support vector machine (SVM):* In machine learning, another common technique that can be used for classification, regression, or other tasks is a support vector machine (SVM) [56]. In high- or infinite-dimensional space, a support vector machine constructs a hyper-plane or set of hyper-planes. Intuitively, the hyper-plane, which has the greatest distance from the nearest training data points in any class, achieves a strong separation since, in general, the greater the margin, the lower the classifier's generalization error. It is effective in high-dimensional spaces and can behave differently based on different mathematical functions known as the kernel. Linear, polynomial, radial basis function (RBF), sigmoid, etc., are the popular kernel functions used in SVM classifier [82]. However, when the data set contains more noise, such as overlapping target classes, SVM does not perform well.
- *Decision tree (DT):* Decision tree (DT) [88] is a well-known non-parametric supervised learning method. DT learning methods are used for both the classification and regression tasks [82]. ID3 [87], C4.5 [88], and CART [20] are well known for DT algorithms. Moreover, recently proposed BehavDT [100], and IntrudTree [97] by Sarker et al. are effective in the relevant application domains, such as user behavior analytics and cybersecurity analytics, respectively. By sorting down the tree from the root to some leaf nodes, as shown in Fig. 4, DT classifies the instances. Instances are classified by checking the attribute defined by that node, starting at the root node of the tree, and then moving down the tree branch corresponding to the attribute value. For splitting, the most popular criteria are "gini" for the Gini impurity and "entropy" for the information gain that can be expressed mathematically as [82].

$$\text{Entropy}: H(x) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \tag{2}$$

$$\text{Gini}(E) = 1 - \sum_{i=1}^{c} p_i^2. \tag{3}$$

- *Random forest (RF):* A random forest classifier [19] is well known as an ensemble classification technique that is used in the field of machine learning and data science in various application areas. This method uses

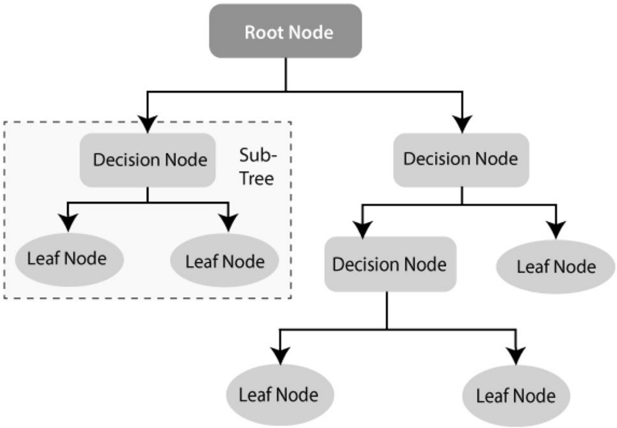


Fig. 4 An example of a decision tree structure

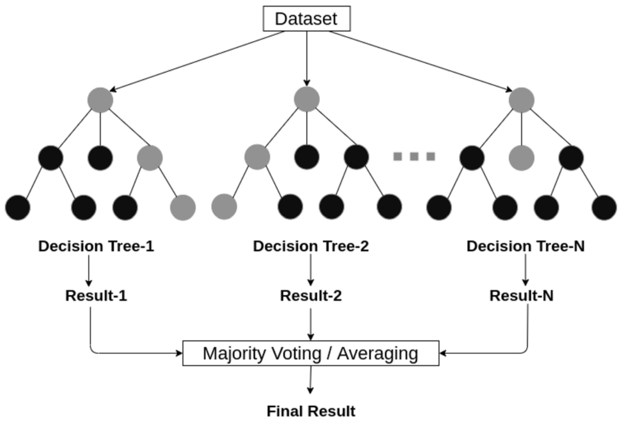


Fig. 5 An example of a random forest structure considering multiple decision trees

"parallel ensembling" which fits several decision tree classifiers in parallel, as shown in Fig. 5, on different data set sub-samples and uses majority voting or averages for the outcome or final result. It thus minimizes the over-fitting problem and increases the prediction accuracy and control [82]. Therefore, the RF learning model with multiple decision trees is typically more accurate than a single decision tree based model [106]. To build a series of decision trees with controlled variation, it combines bootstrap aggregation (bagging) [18] and random feature selection [11]. It is adaptable to both classification and regression problems and fits well for both categorical and continuous values.

- *Adaptive Boosting (AdaBoost):* Adaptive Boosting (AdaBoost) is an ensemble learning process that employs an iterative approach to improve poor classifiers by learning from their errors. This is developed by Yoav Freund et al. [35] and also known as "meta-learning". Unlike the random forest that uses parallel ensembling, Adaboost uses "sequential ensembling". It creates a powerful classifier by combining many poorly performing classifiers to obtain a good classifier of high accuracy. In that sense, AdaBoost is called an adaptive classifier by significantly improving the efficiency of the classifier, but in some instances, it can trigger overfits. AdaBoost is best used to boost the performance of decision trees, base estimator [82], on binary classification problems, however, is sensitive to noisy data and outliers.
- *Extreme gradient boosting (XGBoost):* Gradient Boosting, like Random Forests [19] above, is an ensemble learning algorithm that generates a final model based on a series of individual models, typically decision trees. The gradient is used to minimize the loss function, similar to how neural networks [41] use gradient descent to optimize weights. Extreme Gradient Boosting (XGBoost) is a form of gradient boosting that takes more detailed approximations into account when determining the best model [82]. It computes second-order gradients of the loss function to minimize loss and advanced regularization (L1 and L2) [82], which reduces over-fitting, and improves model generalization and performance. XGBoost is fast to interpret and can handle large-sized datasets well.
- *Stochastic gradient descent (SGD):* Stochastic gradient descent (SGD) [41] is an iterative method for optimizing an objective function with appropriate smoothness properties, where the word 'stochastic' refers to random probability. This reduces the computational burden, particularly in high-dimensional optimization problems, allowing for faster iterations in exchange for a lower convergence rate. A gradient is the slope of a function that calculates a variable's degree of change in response to another variable's changes. Mathematically, the Gradient Descent is a convex function whose output is a partial derivative of a set of its input parameters. Let, α is the learning rate, and J_i is the training example cost of ith, then Eq. (4) represents the stochastic gradient descent weight update method at the j^{th} iteration. In large-scale and sparse machine learning, SGD has been successfully applied to problems often encountered in text classification and natural language processing [82]. However, SGD is sensitive to feature scaling and needs a range of hyperparameters, such as the regularization parameter and the number of iterations.

$$w_j \ := \ w_j - \alpha \frac{\partial J_i}{\partial w_j}. \tag{4}$$

- *Rule-based classification*: The term rule-based classification can be used to refer to any classification scheme that makes use of IF-THEN rules for class prediction. Several classification algorithms such as Zero-R [125], One-R [47], decision trees [87, 88], DTNB [110], Ripple Down Rule learner (RIDOR) [125], Repeated Incremental Pruning to Produce Error Reduction (RIPPER) [126] exist with the ability of rule generation. The decision tree is one of the most common rule-based classification algorithms among these techniques because it has several advantages, such as being easier to interpret; the ability to handle high-dimensional data; simplicity and speed; good accuracy; and the capability to produce rules for human clear and understandable classification [127] [128]. The decision tree-based rules also provide significant accuracy in a prediction model for unseen test cases [106]. Since the rules are easily interpretable, these rule-based classifiers are often used to produce descriptive models that can describe a system including the entities and their relationships.

Regression Analysis

Regression analysis includes several methods of machine learning that allow to predict a continuous (y) result variable based on the value of one or more (x) predictor variables [41]. The most significant distinction between classification and regression is that classification predicts distinct class labels, while regression facilitates the prediction of a continuous quantity. Figure 6 shows an example of how classification is different with regression models. Some overlaps are often found between the two types of machine learning algorithms. Regression models are now widely used in a variety of fields, including financial forecasting or prediction, cost estimation, trend analysis, marketing, time series estimation, drug response modeling, and many more. Some of the familiar types of regression algorithms are linear, polynomial, lasso and ridge regression, etc., which are explained briefly in the following.

- *Simple and multiple linear regression:* This is one of the most popular ML modeling techniques as well as a well-known regression technique. In this technique, the dependent variable is continuous, the independent variable(s) can be continuous or discrete, and the form of the regression line is linear. Linear regression creates a relationship between the dependent variable (Y) and one or more independent variables (X) (also known as regression line) using the best fit straight line [41]. It is defined by the following equations:

$$y = a + bx + e \tag{5}$$

$$y = a + b_1x_1 + b_2x_2 + \cdots + b_nx_n + e, \tag{6}$$

where a is the intercept, b is the slope of the line, and e is the error term. This equation can be used to predict the value of the target variable based on the given predictor variable(s). Multiple linear regression is an extension of simple linear regression that allows two or more predictor variables to model a response variable, y, as a linear function [41] defined in Eq. 6, whereas simple linear regression has only 1 independent variable, defined in Eq. 5.

- *Polynomial regression:* Polynomial regression is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is not linear, but is the polynomial degree of n^{th} in x [82]. The equation for polynomial regression is also derived from linear regression (polynomial regression of degree 1) equation, which is defined as below:

$$y = b_0 + b_1x + b_2x^2 + b_3x^3 + \cdots + b_nx^n + e. \tag{7}$$

Here, y is the predicted/target output, $b_0, b_1, ...b_n$ are the regression coefficients, x is an independent/ input variable. In simple words, we can say that if data are not distributed linearly, instead it is n^{th} degree of polynomial then we use polynomial regression to get desired output.

- *LASSO and ridge regression:* LASSO and Ridge regression are well known as powerful techniques which are typically used for building learning models in presence of a large number of features, due to their capability to preventing over-fitting and reducing the complexity of the model. The LASSO (least absolute shrinkage and selection operator) regression model uses $L1$ regularization technique [82] that uses shrinkage, which penalizes "absolute value of magnitude of coefficients" ($L1$ penalty). As a result, LASSO appears to render coefficients to absolute zero. Thus, LASSO regression aims to find the subset of predictors that minimizes the prediction error for a quantitative response variable. On the other hand, ridge regression uses $L2$ regularization [82], which is the "squared magnitude of coefficients" ($L2$ penalty). Thus, ridge regression forces the weights to be small but never sets the coefficient value to zero, and does a non-sparse solution. Overall, LASSO regression is useful to obtain a subset of predictors by eliminating less important features, and ridge regression is useful when a data set has "multicollinearity" which refers to the predictors that are correlated with other predictors.

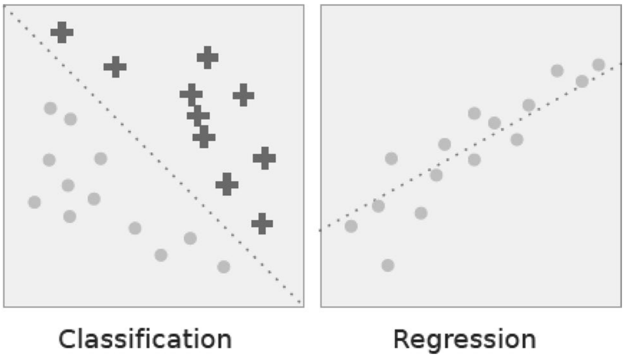


Fig. 6 Classification vs. regression. In classification the dotted line represents a linear boundary that separates the two classes; in regression, the dotted line models the linear relationship between the two variables

Cluster Analysis

Cluster analysis, also known as clustering, is an unsupervised machine learning technique for identifying and grouping related data points in large datasets without concern for the specific outcome. It does grouping a collection of objects in such a way that objects in the same category, called a cluster, are in some sense more similar to each other than objects in other groups [41]. It is often used as a data analysis technique to discover interesting trends or patterns in data, e.g., groups of consumers based on their behavior. In a broad range of application areas, such as cybersecurity, e-commerce, mobile data processing, health analytics, user modeling and behavioral analytics, clustering can be used. In the following, we briefly discuss and summarize various types of clustering methods.

- *Partitioning methods:* Based on the features and similarities in the data, this clustering approach categorizes the data into multiple groups or clusters. The data scientists or analysts typically determine the number of clusters either dynamically or statically depending on the nature of the target applications, to produce for the methods of clustering. The most common clustering algorithms based on partitioning methods are K-means [69], K-Mediods [80], CLARA [55] etc.
- *Density-based methods:* To identify distinct groups or clusters, it uses the concept that a cluster in the data space is a contiguous region of high point density isolated from other such clusters by contiguous regions of low point density. Points that are not part of a cluster are considered as noise. The typical clustering algorithms based on density are DBSCAN [32], OPTICS [12] etc. The density-based methods typically struggle with clusters of similar density and high dimensionality data.
- *Hierarchical-based methods:* Hierarchical clustering typically seeks to construct a hierarchy of clusters, i.e., the tree structure. Strategies for hierarchical clustering generally fall into two types: (i) Agglomerative—a "bottom-up" approach in which each observation begins in its cluster and pairs of clusters are combined as one, moves up the hierarchy, and (ii) Divisive—a "top-down" approach in which all observations begin in one cluster and splits are performed recursively, moves down the hierarchy, as shown in Fig 7. Our earlier proposed BOTS technique, Sarker et al. [102] is an example of a hierarchical, particularly, bottom-up clustering algorithm.
- *Grid-based methods:* To deal with massive datasets, grid-based clustering is especially suitable. To obtain clusters, the principle is first to summarize the dataset with a grid representation and then to combine grid cells. STING [122], CLIQUE [6], etc. are the standard algorithms of grid-based clustering.
- *Model-based methods:* There are mainly two types of model-based clustering algorithms: one that uses statistical learning, and the other based on a method of neural network learning [130]. For instance, GMM [89] is an example of a statistical learning method, and SOM [22] [96] is an example of a neural network learning method.
- *Constraint-based methods:* Constrained-based clustering is a semi-supervised approach to data clustering that uses constraints to incorporate domain knowledge. Application or user-oriented constraints are incorporated to perform the clustering. The typical algorithms of this kind of clustering are COP K-means [121], CMWK-Means [27], etc.

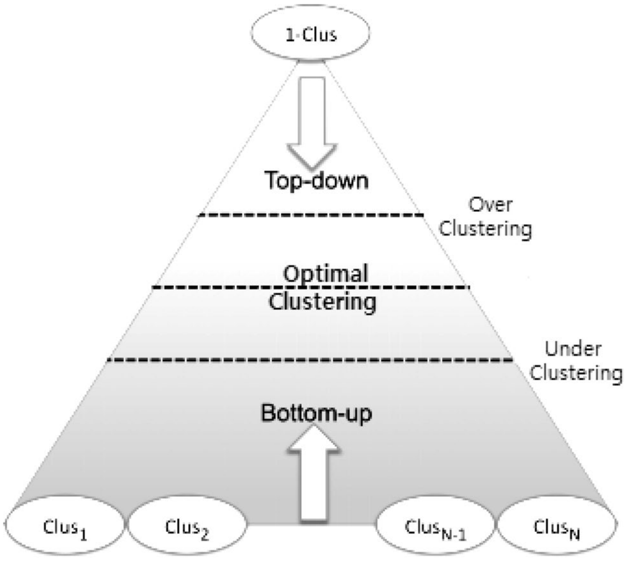


Fig. 7 A graphical interpretation of the widely-used hierarchical clustering (Bottom-up and top-down) technique

Many clustering algorithms have been proposed with the ability to grouping data in machine learning and data science literature [41, 125]. In the following, we summarize the popular methods that are used widely in various application areas.

- *K-means clustering:* K-means clustering [69] is a fast, robust, and simple algorithm that provides reliable results when data sets are well-separated from each other. The data points are allocated to a cluster in this algorithm in such a way that the amount of the squared distance between the data points and the centroid is as small as possible. In other words, the K-means algorithm identifies the *k* number of centroids and then assigns each data point to the nearest cluster while keeping the centroids as small as possible. Since it begins with a random selection of cluster centers, the results can be inconsistent. Since extreme values can easily affect a mean, the K-means clustering algorithm is sensitive to outliers. K-medoids clustering [91] is a variant of K-means that is more robust to noises and outliers.
- *Mean-shift clustering:* Mean-shift clustering [37] is a nonparametric clustering technique that does not require prior knowledge of the number of clusters or constraints on cluster shape. Mean-shift clustering aims to discover "blobs" in a smooth distribution or

density of samples [82]. It is a centroid-based algorithm that works by updating centroid candidates to be the mean of the points in a given region. To form the final set of centroids, these candidates are filtered in a post-processing stage to remove near-duplicates. Cluster analysis in computer vision and image processing are examples of application domains. Mean Shift has the disadvantage of being computationally expensive. Moreover, in cases of high dimension, where the number of clusters shifts abruptly, the mean-shift algorithm does not work well.

- *DBSCAN:* Density-based spatial clustering of applications with noise (DBSCAN) [32] is a base algorithm for density-based clustering which is widely used in data mining and machine learning. This is known as a non-parametric density-based clustering technique for separating high-density clusters from low-density clusters that are used in model building. DBSCAN's main idea is that a point belongs to a cluster if it is close to many points from that cluster. It can find clusters of various shapes and sizes in a vast volume of data that is noisy and contains outliers. DBSCAN, unlike k-means, does not require a priori specification of the number of clusters in the data and can find arbitrarily shaped clusters. Although k-means is much faster than DBSCAN, it is efficient at finding high-density regions and outliers, i.e., is robust to outliers.
- *GMM clustering:* Gaussian mixture models (GMMs) are often used for data clustering, which is a distribution-based clustering algorithm. A Gaussian mixture model is a probabilistic model in which all the data points are produced by a mixture of a finite number of Gaussian distributions with unknown parameters [82]. To find the Gaussian parameters for each cluster, an optimization algorithm called expectation-maximization (EM) [82] can be used. EM is an iterative method that uses a statistical model to estimate the parameters. In contrast to k-means, Gaussian mixture models account for uncertainty and return the likelihood that a data point belongs to one of the k clusters. GMM clustering is more robust than k-means and works well even with non-linear data distributions.
- *Agglomerative hierarchical clustering:* The most common method of hierarchical clustering used to group objects in clusters based on their similarity is agglomerative clustering. This technique uses a bottom-up approach, where each object is first treated as a singleton cluster by the algorithm. Following that, pairs of clusters are merged one by one until all clusters have been merged into a single large cluster containing all objects. The result is a dendrogram, which is a tree-based representation of the elements. Single linkage [115], Complete linkage [116], BOTS [102] etc. are some examples of such techniques. The main advantage of agglomerative hierarchical clustering over k-means is that the tree-structure hierarchy generated by agglomerative clustering is more informative than the unstructured collection of flat clusters returned by k-means, which can help to make better decisions in the relevant application areas.

Dimensionality Reduction and Feature Learning

In machine learning and data science, high-dimensional data processing is a challenging task for both researchers and application developers. Thus, dimensionality reduction which is an unsupervised learning technique, is important because it leads to better human interpretations, lower computational costs, and avoids overfitting and redundancy by simplifying models. Both the process of feature selection and feature extraction can be used for dimensionality reduction. The primary distinction between the selection and extraction of features is that the "feature selection" keeps a subset of the original features [97], while "feature extraction" creates brand new ones [98]. In the following, we briefly discuss these techniques.

- *Feature selection:* The selection of features, also known as the selection of variables or attributes in the data, is the process of choosing a subset of unique features (variables, predictors) to use in building machine learning and data science model. It decreases a model's complexity by eliminating the irrelevant or less important features and allows for faster training of machine learning algorithms. A right and optimal subset of the selected features in a problem domain is capable to minimize the overfitting problem through simplifying and generalizing the model as well as increases the model's accuracy [97]. Thus, "feature selection" [66, 99] is considered as one of the primary concepts in machine learning that greatly affects the effectiveness and efficiency of the target machine learning model. Chi-squared test, Analysis of variance (ANOVA) test, Pearson's correlation coefficient, recursive feature elimination, are some popular techniques that can be used for feature selection.
- *Feature extraction:* In a machine learning-based model or system, feature extraction techniques usually provide a better understanding of the data, a way to improve prediction accuracy, and to reduce computational cost or training time. The aim of "feature extraction" [66, 99] is to reduce the number of features in a dataset by generating new ones from the existing ones and then discarding the original features. The majority of the information found in the original set of features can then be summarized using this new reduced set of features. For instance, principal components analysis (PCA) is often used as a dimensionality-reduction technique to extract a lower-

dimensional space creating new brand components from the existing features in a dataset [98].

Many algorithms have been proposed to reduce data dimensions in the machine learning and data science literature [41, 125]. In the following, we summarize the popular methods that are used widely in various application areas.

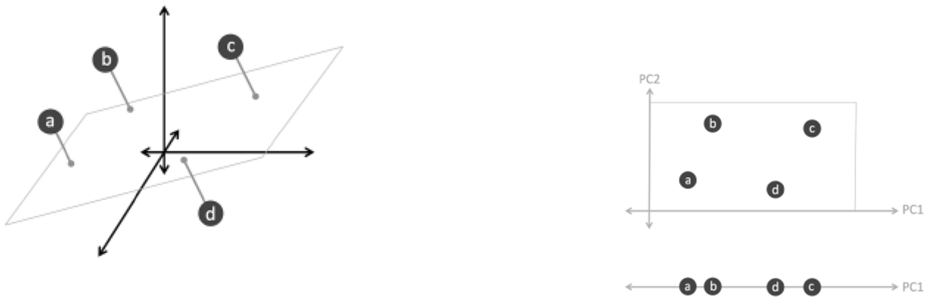
- *Variance threshold:* A simple basic approach to feature selection is the variance threshold [82]. This excludes all features of low variance, i.e., all features whose variance does not exceed the threshold. It eliminates all zero-variance characteristics by default, i.e., characteristics that have the same value in all samples. This feature selection algorithm looks only at the (X) features, not the (y) outputs needed, and can, therefore, be used for unsupervised learning.
- *Pearson correlation:* Pearson's correlation is another method to understand a feature's relation to the response variable and can be used for feature selection [99]. This method is also used for finding the association between the features in a dataset. The resulting value is $[-1, 1]$, where -1 means perfect negative correlation, $+1$ means perfect positive correlation, and 0 means that the two variables do not have a linear correlation. If two random variables represent X and Y, then the correlation coefficient between X and Y is defined as [41]

$$r(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}. \tag{8}$$

- *ANOVA:* Analysis of variance (ANOVA) is a statistical tool used to verify the mean values of two or more groups that differ significantly from each other. ANOVA assumes a linear relationship between the variables and the target and the variables' normal distribution. To statistically test the equality of means, the ANOVA method utilizes F tests. For feature selection, the results 'ANOVA F value' [82] of this test can be used where certain features independent of the goal variable can be omitted.
- *Chi square:* The chi-square χ^2 [82] statistic is an estimate of the difference between the effects of a series of events or variables observed and expected frequencies. The magnitude of the difference between the real and observed values, the degrees of freedom, and the sample size depends on χ^2. The chi-square χ^2 is commonly used for testing relationships between categorical variables. If O_i represents observed value and E_i represents expected value, then

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}. \tag{9}$$

- *Recursive feature elimination (RFE):* Recursive Feature Elimination (RFE) is a brute force approach to feature selection. RFE [82] fits the model and removes the weakest feature before it meets the specified number of features. Features are ranked by the coefficients or feature significance of the model. RFE aims to remove dependencies and collinearity in the model by recursively removing a small number of features per iteration.
- *Model-based selection:* To reduce the dimensionality of the data, linear models penalized with the $L1$ regularization can be used. Least absolute shrinkage and selection operator (Lasso) regression is a type of linear regression that has the property of shrinking some of the coefficients to zero [82]. Therefore, that feature can be removed from the model. Thus, the penalized lasso regression method, often used in machine learning to select the subset of variables. Extra Trees Classifier [82] is an example of a tree-based estimator that can be used to compute impurity-based function importance, which can then be used to discard irrelevant features.
- *Principal component analysis (PCA):* Principal component analysis (PCA) is a well-known unsupervised learn-



(a) An example of the original features in a 3D space. (b) Principal components in 2D and 1D space.

Fig. 8 An example of a principal component analysis (PCA) and created principal components PC1 and PC2 in different dimension space

ing approach in the field of machine learning and data science. PCA is a mathematical technique that transforms a set of correlated variables into a set of uncorrelated variables known as principal components [48, 81]. Figure 8 shows an example of the effect of PCA on various dimensions space, where Fig. 8a shows the original features in 3D space, and Fig. 8b shows the created principal components PC1 and PC2 onto a 2D plane, and 1D line with the principal component PC1 respectively. Thus, PCA can be used as a feature extraction technique that reduces the dimensionality of the datasets, and to build an effective machine learning model [98]. Technically, PCA identifies the completely transformed with the highest eigenvalues of a covariance matrix and then uses those to project the data into a new subspace of equal or fewer dimensions [82].

Association Rule Learning

Association rule learning is a rule-based machine learning approach to discover interesting relationships, "IF-THEN" statements, in large datasets between variables [7]. One example is that "if a customer buys a computer or laptop (an item), s/he is likely to also buy anti-virus software (another item) at the same time". Association rules are employed today in many application areas, including IoT services, medical diagnosis, usage behavior analytics, web usage mining, smartphone applications, cybersecurity applications, and bioinformatics. In comparison to sequence mining, association rule learning does not usually take into account the order of things within or across transactions. A common way of measuring the usefulness of association rules is to use its parameter, the 'support' and 'confidence', which is introduced in [7].

In the data mining literature, many association rule learning methods have been proposed, such as logic dependent [34], frequent pattern based [8, 49, 68], and tree-based [42]. The most popular association rule learning algorithms are summarized below.

- *AIS and SETM:* AIS is the first algorithm proposed by Agrawal et al. [7] for association rule mining. The AIS algorithm's main downside is that too many candidate itemsets are generated, requiring more space and wasting a lot of effort. This algorithm calls for too many passes over the entire dataset to produce the rules. Another approach SETM [49] exhibits good performance and stable behavior with execution time; however, it suffers from the same flaw as the AIS algorithm.
- *Apriori:* For generating association rules for a given dataset, Agrawal et al. [8] proposed the Apriori, Apriori-TID, and Apriori-Hybrid algorithms. These later algorithms outperform the AIS and SETM mentioned above due to the Apriori property of frequent itemset [8]. The term 'Apriori' usually refers to having prior knowledge of frequent itemset properties. Apriori uses a "bottom-up" approach, where it generates the candidate itemsets. To reduce the search space, Apriori uses the property "all subsets of a frequent itemset must be frequent; and if an itemset is infrequent, then all its supersets must also be infrequent". Another approach predictive Apriori [108] can also generate rules; however, it receives unexpected results as it combines both the support and confidence. The Apriori [8] is the widely applicable techniques in mining association rules.
- *ECLAT:* This technique was proposed by Zaki et al. [131] and stands for Equivalence Class Clustering and bottom-up Lattice Traversal. ECLAT uses a depth-first search to find frequent itemsets. In contrast to the Apriori [8] algorithm, which represents data in a horizontal pattern, it represents data vertically. Hence, the ECLAT algorithm is more efficient and scalable in the area of association rule learning. This algorithm is better suited for small and medium datasets whereas the Apriori algorithm is used for large datasets.
- *FP-Growth:* Another common association rule learning technique based on the frequent-pattern tree (FP-tree) proposed by Han et al. [42] is Frequent Pattern Growth, known as FP-Growth. The key difference with Apriori is that while generating rules, the Apriori algorithm [8] generates frequent candidate itemsets; on the other hand, the FP-growth algorithm [42] prevents candidate generation and thus produces a tree by the successful strategy of 'divide and conquer' approach. Due to its sophistication, however, FP-Tree is challenging to use in an interactive mining environment [133]. Thus, the FP-Tree would not fit into memory for massive data sets, making it challenging to process big data as well. Another solution is RARM (Rapid Association Rule Mining) proposed by Das et al. [26] but faces a related FP-tree issue [133].
- *ABC-RuleMiner:* A rule-based machine learning method, recently proposed in our earlier paper, by Sarker et al. [104], to discover the interesting non-redundant rules to provide real-world intelligent services. This algorithm effectively identifies the redundancy in associations by taking into account the impact or precedence of the related contextual features and discovers a set of non-redundant association rules. This algorithm first constructs an association generation tree (AGT), a top-down approach, and then extracts the association rules through traversing the tree. Thus, ABC-RuleMiner is more potent than traditional rule-based methods in terms of both non-redundant rule generation and intelligent decision-making, particularly in a context-aware smart computing environment, where human or user preferences are involved.

Among the association rule learning techniques discussed above, Apriori [8] is the most widely used algorithm for discovering association rules from a given dataset [133]. The main strength of the association learning technique is its comprehensiveness, as it generates all associations that satisfy the user-specified constraints, such as minimum support and confidence value. The ABC-RuleMiner approach [104] discussed earlier could give significant results in terms of non-redundant rule generation and intelligent decision-making for the relevant application areas in the real world.

Reinforcement Learning

Reinforcement learning (RL) is a machine learning technique that allows an agent to learn by trial and error in an interactive environment using input from its actions and experiences. Unlike supervised learning, which is based on given sample data or examples, the RL method is based on interacting with the environment. The problem to be solved in reinforcement learning (RL) is defined as a Markov Decision Process (MDP) [86], i.e., all about sequentially making decisions. An RL problem typically includes four elements such as Agent, Environment, Rewards, and Policy.

RL can be split roughly into Model-based and Model-free techniques. Model-based RL is the process of inferring optimal behavior from a model of the environment by performing actions and observing the results, which include the next state and the immediate reward [85]. AlphaZero, AlphaGo [113] are examples of the model-based approaches. On the other hand, a model-free approach does not use the distribution of the transition probability and the reward function associated with MDP. Q-learning, Deep Q Network, Monte Carlo Control, SARSA (State–Reward–State–Action), etc. are some examples of model-free algorithms [52]. The policy network, which is required for model-based RL but not for model-free, is the key difference between model-free and model-based learning. In the following, we discuss the popular RL algorithms.

- *Monte Carlo methods:* Monte Carlo techniques, or Monte Carlo experiments, are a wide category of computational algorithms that rely on repeated random sampling to obtain numerical results [52]. The underlying concept is to use randomness to solve problems that are deterministic in principle. Optimization, numerical integration, and making drawings from the probability distribution are the three problem classes where Monte Carlo techniques are most commonly used.
- *Q-learning:* Q-learning is a model-free reinforcement learning algorithm for learning the quality of behaviors that tell an agent what action to take under what conditions [52]. It does not need a model of the environment (hence the term “model-free”), and it can deal with stochastic transitions and rewards without the need for adaptations. The ‘Q’ in Q-learning usually stands for quality, as the algorithm calculates the maximum expected rewards for a given behavior in a given state.
- *Deep Q-learning:* The basic working step in Deep Q-Learning [52] is that the initial state is fed into the neural network, which returns the Q-value of all possible actions as an output. Still, when we have a reasonably simple setting to overcome, Q-learning works well. However, when the number of states and actions becomes more complicated, deep learning can be used as a function approximator.

Reinforcement learning, along with supervised and unsupervised learning, is one of the basic machine learning paradigms. RL can be used to solve numerous real-world problems in various fields, such as game theory, control theory, operations analysis, information theory, simulation-based optimization, manufacturing, supply chain logistics, multi-agent systems, swarm intelligence, aircraft control, robot motion control, and many more.

Artificial Neural Network and Deep Learning

Deep learning is part of a wider family of artificial neural networks (ANN)-based machine learning approaches with representation learning. Deep learning provides a computational architecture by combining several processing layers, such as input, hidden, and output layers, to learn from data [41]. The main advantage of deep learning over traditional machine learning methods is its better performance in several cases, particularly learning from large datasets [105, 129]. Figure 9 shows a general performance of deep learning over machine learning considering the increasing amount of data. However, it may vary depending on the data characteristics and experimental set up.

The most common deep learning algorithms are: Multi-layer Perceptron (MLP), Convolutional Neural Network

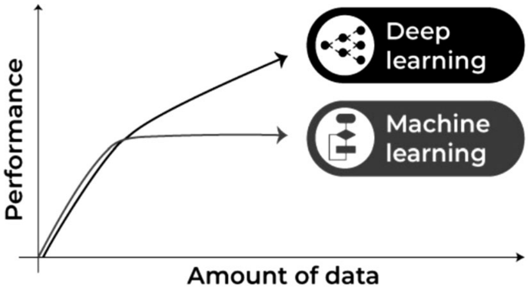


Fig. 9 Machine learning and deep learning performance in general with the amount of data

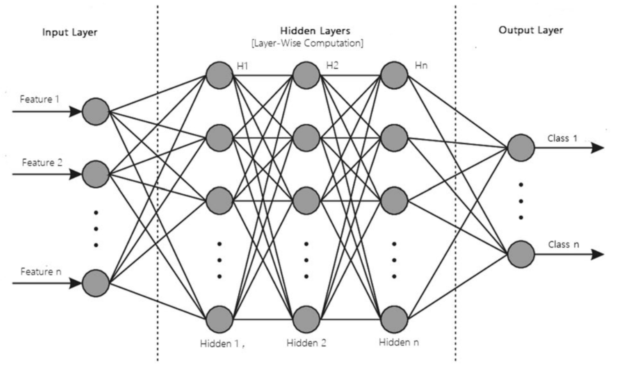


Fig. 10 A structure of an artificial neural network modeling with multiple processing layers

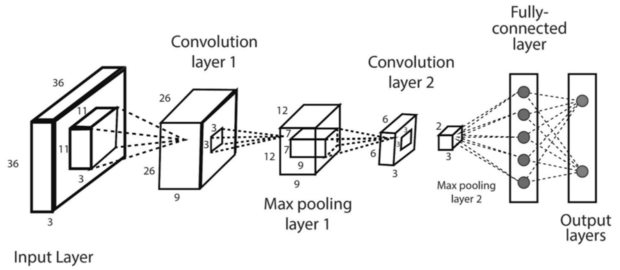


Fig. 11 An example of a convolutional neural network (CNN or ConvNet) including multiple convolution and pooling layers

(CNN, or ConvNet), Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) [96]. In the following, we discuss various types of deep learning methods that can be used to build effective data-driven models for various purposes.

- *MLP:* The base architecture of deep learning, which is also known as the feed-forward artificial neural network, is called a multilayer perceptron (MLP) [82]. A typical MLP is a fully connected network consisting of an input layer, one or more hidden layers, and an output layer, as shown in Fig. 10. Each node in one layer connects to each node in the following layer at a certain weight. MLP utilizes the "Backpropagation" technique [41], the most "fundamental building block" in a neural network, to adjust the weight values internally while building the model. MLP is sensitive to scaling features and allows a variety of hyperparameters to be tuned, such as the number of hidden layers, neurons, and iterations, which can result in a computationally costly model.
- *CNN or ConvNet:* The convolution neural network (CNN) [65] enhances the design of the standard ANN, consisting of convolutional layers, pooling layers, as well as fully connected layers, as shown in Fig. 11. As it takes the advantage of the two-dimensional (2D) structure of the input data, it is typically broadly used in several areas such as image and video recognition, image processing and classification, medical image analysis, natural language processing, etc. While CNN has a greater computational burden, without any manual intervention, it has the advantage of automatically detecting the important features, and hence CNN is considered to be more powerful than conventional ANN. A number of advanced deep learning models based on CNN can be used in the field, such as AlexNet [60], Xception [24], Inception [118], Visual Geometry Group (VGG) [44], ResNet [45], etc.
- *LSTM-RNN:* Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the area of deep learning [38]. LSTM has feedback links, unlike normal feed-forward neural networks. LSTM networks are well-suited for analyzing and learning sequential data, such as classifying, processing, and predicting data based on time series data, which differentiates it from other conventional networks. Thus, LSTM can be used when the data are in a sequential format, such as time, sentence, etc., and commonly applied in the area of time-series analysis, natural language processing, speech recognition, etc.

In addition to these most common deep learning methods discussed above, several other deep learning approaches [96] exist in the area for various purposes. For instance, the self-organizing map (SOM) [58] uses unsupervised learning to represent the high-dimensional data by a 2D grid map, thus achieving dimensionality reduction. The autoencoder (AE) [15] is another learning technique that is widely used for dimensionality reduction as well and feature extraction in unsupervised learning tasks. Restricted Boltzmann machines (RBM) [46] can be used for dimensionality reduction, classification, regression, collaborative filtering, feature learning, and topic modeling. A deep belief network (DBN) is typically composed of simple, unsupervised networks such as restricted Boltzmann machines (RBMs) or autoencoders, and a backpropagation neural network (BPNN) [123]. A generative adversarial network (GAN) [39] is a form of the network for deep learning that can generate data with characteristics close to the actual data input. Transfer learning is currently very common because it can train deep neural networks with comparatively low data, which is typically the re-use of a new problem with a pre-trained model [124]. A brief discussion of these artificial neural networks (ANN) and deep learning (DL) models are summarized in our earlier paper Sarker et al. [96].

Overall, based on the learning techniques discussed above, we can conclude that various types of machine learning techniques, such as classification analysis, regression, data clustering, feature selection and extraction, and dimensionality reduction, association rule learning, reinforcement learning, or deep learning techniques, can play a significant role for various purposes according to their capabilities. In the following section, we discuss several application areas based on machine learning algorithms.

Applications of Machine Learning

In the current age of the Fourth Industrial Revolution (4IR), machine learning becomes popular in various application areas, because of its learning capabilities from the past and making intelligent decisions. In the following, we summarize and discuss ten popular application areas of machine learning technology.

- *Predictive analytics and intelligent decision-making:* A major application field of machine learning is intelligent decision-making by data-driven predictive analytics [21, 70]. The basis of predictive analytics is capturing and exploiting relationships between explanatory variables and predicted variables from previous events to predict the unknown outcome [41]. For instance, identifying suspects or criminals after a crime has been committed, or detecting credit card fraud as it happens. Another application, where machine learning algorithms can assist retailers in better understanding consumer preferences and behavior, better manage inventory, avoiding out-of-stock situations, and optimizing logistics and warehousing in e-commerce. Various machine learning algorithms such as decision trees, support vector machines, artificial neural networks, etc. [106, 125] are commonly used in the area. Since accurate predictions provide insight into the unknown, they can improve the decisions of industries, businesses, and almost any organization, including government agencies, e-commerce, telecommunications, banking and financial services, healthcare, sales and marketing, transportation, social networking, and many others.
- *Cybersecurity and threat intelligence:* Cybersecurity is one of the most essential areas of Industry 4.0. [114], which is typically the practice of protecting networks, systems, hardware, and data from digital attacks [114]. Machine learning has become a crucial cybersecurity technology that constantly learns by analyzing data to identify patterns, better detect malware in encrypted traffic, find insider threats, predict where bad neighborhoods are online, keep people safe while browsing, or secure data in the cloud by uncovering suspicious activity. For instance, clustering techniques can be used to identify cyber-anomalies, policy violations, etc. To detect various types of cyber-attacks or intrusions machine learning classification models by taking into account the impact of security features are useful [97]. Various deep learning-based security models can also be used on the large scale of security datasets [96, 129]. Moreover, security policy rules generated by association rule learning techniques can play a significant role to build a rule-based security system [105]. Thus, we can say that various learning techniques discussed in Sect. Machine Learning Tasks and Algorithms, can enable cybersecurity professionals to be more proactive inefficiently preventing threats and cyber-attacks.
- *Internet of things (IoT) and smart cities:* Internet of Things (IoT) is another essential area of Industry 4.0. [114], which turns everyday objects into smart objects by allowing them to transmit data and automate tasks without the need for human interaction. IoT is, therefore, considered to be the big frontier that can enhance almost all activities in our lives, such as smart governance, smart home, education, communication, transportation, retail, agriculture, health care, business, and many more [70]. Smart city is one of IoT's core fields of application, using technologies to enhance city services and residents' living experiences [132, 135]. As machine learning utilizes experience to recognize trends and create models that help predict future behavior and events, it has become a crucial technology for IoT applications [103]. For example, to predict traffic in smart cities, parking availability prediction, estimate the total usage of energy of the citizens for a particular period, make context-aware and timely decisions for the people, etc. are some tasks that can be solved using machine learning techniques according to the current needs of the people.
- *Traffic prediction and transportation:* Transportation systems have become a crucial component of every country's economic development. Nonetheless, several cities around the world are experiencing an excessive rise in traffic volume, resulting in serious issues such as delays, traffic congestion, higher fuel prices, increased CO_2 pollution, accidents, emergencies, and a decline in modern society's quality of life [40]. Thus, an intelligent transportation system through predicting future traffic is important, which is an indispensable part of a smart city. Accurate traffic prediction based on machine and deep learning modeling can help to minimize the issues [17, 30, 31]. For example, based on the travel history and trend of traveling through various routes, machine learning can assist transportation companies in predicting possible issues that may occur on specific routes and recommending their customers to take a different path. Ultimately, these learning-based data-driven models help improve traffic flow, increase the usage and efficiency of

sustainable modes of transportation, and limit real-world disruption by modeling and visualizing future changes.

- *Healthcare and COVID-19 pandemic:* Machine learning can help to solve diagnostic and prognostic problems in a variety of medical domains, such as disease prediction, medical knowledge extraction, detecting regularities in data, patient management, etc. [33, 77, 112]. Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus, according to the World Health Organization (WHO) [3]. Recently, the learning techniques have become popular in the battle against COVID-19 [61, 63]. For the COVID-19 pandemic, the learning techniques are used to classify patients at high risk, their mortality rate, and other anomalies [61]. It can also be used to better understand the virus's origin, COVID-19 outbreak prediction, as well as for disease diagnosis and treatment [14, 50]. With the help of machine learning, researchers can forecast where and when, the COVID-19 is likely to spread, and notify those regions to match the required arrangements. Deep learning also provides exciting solutions to the problems of medical image processing and is seen as a crucial technique for potential applications, particularly for COVID-19 pandemic [10, 78, 111]. Overall, machine and deep learning techniques can help to fight the COVID-19 virus and the pandemic as well as intelligent clinical decisions making in the domain of healthcare.
- *E-commerce and product recommendations:* Product recommendation is one of the most well known and widely used applications of machine learning, and it is one of the most prominent features of almost any e-commerce website today. Machine learning technology can assist businesses in analyzing their consumers' purchasing histories and making customized product suggestions for their next purchase based on their behavior and preferences. E-commerce companies, for example, can easily position product suggestions and offers by analyzing browsing trends and click-through rates of specific items. Using predictive modeling based on machine learning techniques, many online retailers, such as Amazon [71], can better manage inventory, prevent out-of-stock situations, and optimize logistics and warehousing. The future of sales and marketing is the ability to capture, evaluate, and use consumer data to provide a customized shopping experience. Furthermore, machine learning techniques enable companies to create packages and content that are tailored to the needs of their customers, allowing them to maintain existing customers while attracting new ones.
- *NLP and sentiment analysis:* Natural language processing (NLP) involves the reading and understanding of spoken or written language through the medium of a computer [79, 103]. Thus, NLP helps computers, for instance, to read a text, hear speech, interpret it, analyze sentiment, and decide which aspects are significant, where machine learning techniques can be used. Virtual personal assistant, chatbot, speech recognition, document description, language or machine translation, etc. are some examples of NLP-related tasks. Sentiment Analysis [90] (also referred to as opinion mining or emotion AI) is an NLP sub-field that seeks to identify and extract public mood and views within a given text through blogs, reviews, social media, forums, news, etc. For instance, businesses and brands use sentiment analysis to understand the social sentiment of their brand, product, or service through social media platforms or the web as a whole. Overall, sentiment analysis is considered as a machine learning task that analyzes texts for polarity, such as "positive", "negative", or "neutral" along with more intense emotions like very happy, happy, sad, very sad, angry, have interest, or not interested etc.
- *Image, speech and pattern recognition:* Image recognition [36] is a well-known and widespread example of machine learning in the real world, which can identify an object as a digital image. For instance, to label an x-ray as cancerous or not, character recognition, or face detection in an image, tagging suggestions on social media, e.g., Facebook, are common examples of image recognition. Speech recognition [23] is also very popular that typically uses sound and linguistic models, e.g., Google Assistant, Cortana, Siri, Alexa, etc. [67], where machine learning methods are used. Pattern recognition [13] is defined as the automated recognition of patterns and regularities in data, e.g., image analysis. Several machine learning techniques such as classification, feature selection, clustering, or sequence labeling methods are used in the area.
- *Sustainable agriculture:* Agriculture is essential to the survival of all human activities [109]. Sustainable agriculture practices help to improve agricultural productivity while also reducing negative impacts on the environment [5, 25, 109]. The sustainable agriculture supply chains are knowledge-intensive and based on information, skills, technologies, etc., where knowledge transfer encourages farmers to enhance their decisions to adopt sustainable agriculture practices utilizing the increasing amount of data captured by emerging technologies, e.g., the Internet of Things (IoT), mobile technologies and devices, etc. [5, 53, 54]. Machine learning can be applied in various phases of sustainable agriculture, such as in the pre-production phase - for the prediction of crop yield, soil properties, irrigation requirements, etc.; in the production phase—for weather prediction, disease detection, weed detection, soil nutrient management, livestock management, etc.; in processing phase—for demand estimation, production planning, etc. and in the distribution

phase - the inventory management, consumer analysis, etc.

- *User behavior analytics and context-aware smartphone applications:* Context-awareness is a system's ability to capture knowledge about its surroundings at any moment and modify behaviors accordingly [28, 93]. Context-aware computing uses software and hardware to automatically collect and interpret data for direct responses. The mobile app development environment has been changed greatly with the power of AI, particularly, machine learning techniques through their learning capabilities from contextual data [103, 136]. Thus, the developers of mobile apps can rely on machine learning to create smart apps that can understand human behavior, support, and entertain users [107, 137, 140]. To build various personalized data-driven context-aware systems, such as smart interruption management, smart mobile recommendation, context-aware smart searching, decision-making that intelligently assist end mobile phone users in a pervasive computing environment, machine learning techniques are applicable. For example, context-aware association rules can be used to build an intelligent phone call application [104]. Clustering approaches are useful in capturing users' diverse behavioral activities by taking into account data in time series [102]. To predict the future events in various contexts, the classification methods can be used [106, 139]. Thus, various learning techniques discussed in Sect. "Machine Learning Tasks and Algorithms" can help to build context-aware adaptive and smart applications according to the preferences of the mobile phone users.

In addition to these application areas, machine learning-based models can also apply to several other domains such as bioinformatics, cheminformatics, computer networks, DNA sequence classification, economics and banking, robotics, advanced engineering, and many more.

Challenges and Research Directions

Our study on machine learning algorithms for intelligent data analysis and applications opens several research issues in the area. Thus, in this section, we summarize and discuss the challenges faced and the potential research opportunities and future directions.

In general, the effectiveness and the efficiency of a machine learning-based solution depend on the nature and characteristics of the data, and the performance of the learning algorithms. To collect the data in the relevant domain, such as cybersecurity, IoT, healthcare and agriculture discussed in Sect. "Applications of Machine Learning" is not straightforward, although the current cyberspace enables the production of a huge amount of data with very high frequency. Thus, collecting useful data for the target machine learning-based applications, e.g., smart city applications, and their management is important to further analysis. Therefore, a more in-depth investigation of data collection methods is needed while working on the real-world data. Moreover, the historical data may contain many ambiguous values, missing values, outliers, and meaningless data. The machine learning algorithms, discussed in Sect "Machine Learning Tasks and Algorithms" highly impact on data quality, and availability for training, and consequently on the resultant model. Thus, to accurately clean and pre-process the diverse data collected from diverse sources is a challenging task. Therefore, effectively modifying or enhance existing pre-processing methods, or proposing new data preparation techniques are required to effectively use the learning algorithms in the associated application domain.

To analyze the data and extract insights, there exist many machine learning algorithms, summarized in Sect. "Machine Learning Tasks and Algorithms". Thus, selecting a proper learning algorithm that is suitable for the target application is challenging. The reason is that the outcome of different learning algorithms may vary depending on the data characteristics [106]. Selecting a wrong learning algorithm would result in producing unexpected outcomes that may lead to loss of effort, as well as the model's effectiveness and accuracy. In terms of model building, the techniques discussed in Sect. "Machine Learning Tasks and Algorithms" can directly be used to solve many real-world issues in diverse domains, such as cybersecurity, smart cities and healthcare summarized in Sect. "Applications of Machine Learning". However, the hybrid learning model, e.g., the ensemble of methods, modifying or enhancement of the existing learning techniques, or designing new learning methods, could be a potential future work in the area.

Thus, the ultimate success of a machine learning-based solution and corresponding applications mainly depends on both the data and the learning algorithms. If the data are bad to learn, such as non-representative, poor-quality, irrelevant features, or insufficient quantity for training, then the machine learning models may become useless or will produce lower accuracy. Therefore, effectively processing the data and handling the diverse learning algorithms are important, for a machine learning-based solution and eventually building intelligent applications.

Conclusion

In this paper, we have conducted a comprehensive overview of machine learning algorithms for intelligent data analysis and applications. According to our goal, we have briefly discussed how various types of machine learning methods can be used for making solutions to various real-world issues. A successful machine learning model depends on both the data and the performance of the learning algorithms. The sophisticated learning algorithms then need to be trained through the collected real-world data and knowledge related to the target application before the system can assist with intelligent decision-making. We also discussed several popular application areas based on machine learning techniques to highlight their applicability in various real-world issues. Finally, we have summarized and discussed the challenges faced and the potential research opportunities and future directions in the area. Therefore, the challenges that are identified create promising research opportunities in the field which must be addressed with effective solutions in various application areas. Overall, we believe that our study on machine learning-based solutions opens up a promising direction and can be used as a reference guide for potential research and applications for both academia and industry professionals as well as for decision-makers, from a technical point of view.

Declaration

Conflict of interest The author declares no conflict of interest.

References

1. Canadian institute of cybersecurity, university of new brunswick, iscx dataset, http://www.unb.ca/cic/datasets/index.html/ (Accessed on 20 October 2019).
2. Cic-ddos2019 [online]. available: https://www.unb.ca/cic/datasets/ddos-2019.html/ (Accessed on 28 March 2020).
3. World health organization: WHO. http://www.who.int/.
4. Google trends. In https://trends.google.com/trends/, 2019.
5. Adnan N, Nordin Shahrina Md, Rahman I, Noor A. The effects of knowledge transfer on farmers decision making toward sustainable agriculture practices. World J Sci Technol Sustain Dev. 2018.
6. Agrawal R, Gehrke J, Gunopulos D, Raghavan P. Automatic subspace clustering of high dimensional data for data mining applications. In: Proceedings of the 1998 ACM SIGMOD international conference on Management of data. 1998; 94–105
7. Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. In: ACM SIGMOD Record. ACM. 1993;22: 207–216
8. Agrawal R, Gehrke J, Gunopulos D, Raghavan P. Fast algorithms for mining association rules. In: Proceedings of the International Joint Conference on Very Large Data Bases, Santiago Chile. 1994; 1215: 487–499.
9. Aha DW, Kibler D, Albert M. Instance-based learning algorithms. Mach Learn. 1991;6(1):37–66.
10. Alakus TB, Turkoglu I. Comparison of deep learning approaches to predict covid-19 infection. Chaos Solit Fract. 2020;140:
11. Amit Y, Geman D. Shape quantization and recognition with randomized trees. Neural Comput. 1997;9(7):1545–88.
12. Ankerst M, Breunig MM, Kriegel H-P, Sander J. Optics: ordering points to identify the clustering structure. ACM Sigmod Record. 1999;28(2):49–60.
13. Anzai Y. Pattern recognition and machine learning. Elsevier; 2012.
14. Ardabili SF, Mosavi A, Ghamisi P, Ferdinand F, Varkonyi-Koczy AR, Reuter U, Rabczuk T, Atkinson PM. Covid-19 outbreak prediction with machine learning. Algorithms. 2020;13(10):249.
15. Baldi P. Autoencoders, unsupervised learning, and deep architectures. In: Proceedings of ICML workshop on unsupervised and transfer learning, 2012; 37–49 .
16. Balducci F, Impedovo D, Pirlo G. Machine learning applications on agricultural datasets for smart farm enhancement. Machines. 2018;6(3):38.
17. Boukerche A, Wang J. Machine learning-based traffic prediction models for intelligent transportation systems. Comput Netw. 2020;181
18. Breiman L. Bagging predictors. Mach Learn. 1996;24(2):123–40.
19. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
20. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. CRC Press; 1984.
21. Cao L. Data science: a comprehensive overview. ACM Comput Surv (CSUR). 2017;50(3):43.
22. Carpenter GA, Grossberg S. A massively parallel architecture for a self-organizing neural pattern recognition machine. Comput Vis Graph Image Process. 1987;37(1):54–115.
23. Chiu C-C, Sainath TN, Wu Y, Prabhavalkar R, Nguyen P, Chen Z, Kannan A, Weiss RJ, Rao K, Gonina E, et al. State-of-the-art speech recognition with sequence-to-sequence models. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018 pages 4774–4778. IEEE .
24. Chollet F. Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1251–1258, 2017.
25. Cobuloglu H, Büyüktahtakın IE. A stochastic multi-criteria decision analysis for sustainable biomass crop selection. Expert Syst Appl. 2015;42(15–16):6065–74.
26. Das A, Ng W-K, Woon Y-K. Rapid association rule mining. In: Proceedings of the tenth international conference on Information and knowledge management, pages 474–481. ACM, 2001.
27. de Amorim RC. Constrained clustering with minkowski weighted k-means. In: 2012 IEEE 13th International Symposium on Computational Intelligence and Informatics (CINTI), pages 13–17. IEEE, 2012.
28. Dey AK. Understanding and using context. Person Ubiquit Comput. 2001;5(1):4–7.
29. Eagle N, Pentland AS. Reality mining: sensing complex social systems. Person Ubiquit Comput. 2006;10(4):255–68.
30. Essien A, Petrounias I, Sampaio P, Sampaio S. Improving urban traffic speed prediction using data source fusion and deep learning. In: 2019 IEEE International Conference on Big Data and Smart Computing (BigComp). IEEE. 2019: 1–8. .
31. Essien A, Petrounias I, Sampaio P, Sampaio S. A deep-learning model for urban traffic flow prediction with traffic events mined from twitter. In: World Wide Web, 2020: 1–24 .
32. Ester M, Kriegel H-P, Sander J, Xiaowei X, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd. 1996;96:226–31.
33. Fatima M, Pasha M, et al. Survey of machine learning algorithms for disease diagnostic. J Intell Learn Syst Appl. 2017;9(01):1.
34. Flach PA, Lachiche N. Confirmation-guided discovery of first-order rules with tertius. Mach Learn. 2001;42(1–2):61–95.
35. Freund Y, Schapire RE, et al. Experiments with a new boosting algorithm. In: Icml, Citeseer. 1996; 96: 148–156

36. Fujiyoshi H, Hirakawa T, Yamashita T. Deep learning-based image recognition for autonomous driving. IATSS Res. 2019;43(4):244–52.
37. Fukunaga K, Hostetler L. The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Trans Inform Theory. 1975;21(1):32–40.
38. Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep learning. Cambridge: MIT Press; 2016.
39. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Advances in neural information processing systems. 2014: 2672–2680.
40. Guerrero-Ibáñez J, Zeadally S, Contreras-Castillo J. Sensor technologies for intelligent transportation systems. Sensors. 2018;18(4):1212.
41. Han J, Pei J, Kamber M. Data mining: concepts and techniques. Amsterdam: Elsevier; 2011.
42. Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. In: ACM Sigmod Record, ACM. 2000;29: 1–12.
43. Harmon SA, Sanford TH, Sheng X, Turkbey EB, Roth H, Ziyue X, Yang D, Myronenko A, Anderson V, Amalou A, et al. Artificial intelligence for the detection of covid-19 pneumonia on chest ct using multinational datasets. Nat Commun. 2020;11(1):1–7.
44. He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell. 2015;37(9):1904–16.
45. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 770–778.
46. Hinton GE. A practical guide to training restricted boltzmann machines. In: Neural networks: Tricks of the trade. Springer. 2012; 599-619
47. Holte RC. Very simple classification rules perform well on most commonly used datasets. Mach Learn. 1993;11(1):63–90.
48. Hotelling H. Analysis of a complex of statistical variables into principal components. J Edu Psychol. 1933;24(6):417.
49. Houtsma M, Swami A. Set-oriented mining for association rules in relational databases. In: Data Engineering, 1995. Proceedings of the Eleventh International Conference on, IEEE.1995:25–33.
50. Jamshidi M, Lalbakhsh A, Talla J, Peroutka Z, Hadjilooei F, Lalbakhsh P, Jamshidi M, La Spada L, Mirmozafari M, Dehghani M, et al. Artificial intelligence and covid-19: deep learning approaches for diagnosis and treatment. IEEE Access. 2020;8:109581–95.
51. John GH, Langley P. Estimating continuous distributions in bayesian classifiers. In: Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc. 1995; 338–345
52. Kaelbling LP, Littman ML, Moore AW. Reinforcement learning: a survey. J Artif Intell Res. 1996;4:237–85.
53. Kamble SS, Gunasekaran A, Gawankar SA. Sustainable industry 4.0 framework: a systematic literature review identifying the current trends and future perspectives. Process Saf Environ Protect. 2018;117:408–25.
54. Kamble SS, Gunasekaran A, Gawankar SA. Achieving sustainable performance in a data-driven agriculture supply chain: a review for research and applications. Int J Prod Econ. 2020;219:179–94.
55. Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis, vol. 344. John Wiley & Sons; 2009.
56. Keerthi SS, Shevade SK, Bhattacharyya C, Radha Krishna MK. Improvements to platt's smo algorithm for svm classifier design. Neural Comput. 2001;13(3):637–49.
57. Khadse V, Mahalle PN, Biraris SV. An empirical comparison of supervised machine learning algorithms for internet of things data. In: 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), IEEE. 2018; 1–6
58. Kohonen T. The self-organizing map. Proc IEEE. 1990;78(9):1464–80.
59. Koroniotis N, Moustafa N, Sitnikova E, Turnbull B. Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: bot-iot dataset. Fut Gen Comput Syst. 2019;100:779–96.
60. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, 2012: 1097–1105
61. Kushwaha S, Bahl S, Bagha AK, Parmar KS, Javaid M, Haleem A, Singh RP. Significant applications of machine learning for covid-19 pandemic. J Ind Integr Manag. 2020;5(4).
62. Lade P, Ghosh R, Srinivasan S. Manufacturing analytics and industrial internet of things. IEEE Intell Syst. 2017;32(3):74–9.
63. Lalmuanawma S, Hussain J, Chhakchhuak L. Applications of machine learning and artificial intelligence for covid-19 (sars-cov-2) pandemic: a review. Chaos Sol Fract. 2020:110059 .
64. LeCessie S, Van Houwelingen JC. Ridge estimators in logistic regression. J R Stat Soc Ser C (Appl Stat). 1992;41(1):191–201.
65. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86(11):2278–324.
66. Liu H, Motoda H. Feature extraction, construction and selection: A data mining perspective, vol. 453. Springer Science & Business Media; 1998.
67. López G, Quesada L, Guerrero LA. Alexa vs. siri vs. cortana vs. google assistant: a comparison of speech-based natural user interfaces. In: International Conference on Applied Human Factors and Ergonomics, Springer. 2017; 241–250.
68. Liu B, HsuW, Ma Y. Integrating classification and association rule mining. In: Proceedings of the fourth international conference on knowledge discovery and data mining, 1998.
69. MacQueen J, et al. Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, 1967;volume 1, pages 281–297. Oakland, CA, USA.
70. Mahdavinejad MS, Rezvan M, Barekatain M, Adibi P, Barnaghi P, Sheth AP. Machine learning for internet of things data analysis: a survey. Digit Commun Netw. 2018;4(3):161–75.
71. Marchand A, Marx P. Automated product recommendations with preference-based explanations. J Retail. 2020;96(3):328–43.
72. McCallum A. Information extraction: distilling structured data from unstructured text. Queue. 2005;3(9):48–57.
73. Mehrotra A, Hendley R, Musolesi M. Prefminer: mining user's preferences for intelligent mobile notification management. In: Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing, Heidelberg, Germany, 12–16 September, 2016; pp. 1223–1234. ACM, New York, USA. .
74. Mohamadou Y, Halidou A, Kapen PT. A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of covid-19. Appl Intell. 2020;50(11):3913–25.
75. Mohammed M, Khan MB, Bashier Mohammed BE. Machine learning: algorithms and applications. CRC Press; 2016.
76. Moustafa N, Slay J. Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set). In: 2015 military communications and information systems conference (MilCIS), 2015;pages 1–6. IEEE .
77. Nilashi M, Ibrahim OB, Ahmadi H, Shahmoradi L. An analytical method for diseases prediction using machine learning techniques. Comput Chem Eng. 2017;106:212–23.

78. Yujin O, Park S, Ye JC. Deep learning covid-19 features on cxr using limited training data sets. IEEE Trans Med Imaging. 2020;39(8):2688–700.
79. Otter DW, Medina JR , Kalita JK. A survey of the usages of deep learning for natural language processing. IEEE Trans Neural Netw Learn Syst. 2020.
80. Park H-S, Jun C-H. A simple and fast algorithm for k-medoids clustering. Expert Syst Appl. 2009;36(2):3336–41.
81. Liii Pearson K. on lines and planes of closest fit to systems of points in space. Lond Edinb Dublin Philos Mag J Sci. 1901;2(11):559–72.
82. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–30.
83. Perveen S, Shahbaz M, Keshavjee K, Guergachi A. Metabolic syndrome and development of diabetes mellitus: predictive modeling based on machine learning techniques. IEEE Access. 2018;7:1365–75.
84. Santi P, Ram D, Rob C, Nathan E. Behavior-based adaptive call predictor. ACM Trans Auton Adapt Syst. 2011;6(3):21:1–21:28.
85. Polydoros AS, Nalpantidis L. Survey of model-based reinforcement learning: applications on robotics. J Intell Robot Syst. 2017;86(2):153–73.
86. Puterman ML. Markov decision processes: discrete stochastic dynamic programming. John Wiley & Sons; 2014.
87. Quinlan JR. Induction of decision trees. Mach Learn. 1986;1:81–106.
88. Quinlan JR. C4.5: programs for machine learning. Mach Learn. 1993.
89. Rasmussen C. The infinite gaussian mixture model. Adv Neural Inform Process Syst. 1999;12:554–60.
90. Ravi K, Ravi V. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowl Syst. 2015;89:14–46.
91. Rokach L. A survey of clustering algorithms. In: Data mining and knowledge discovery handbook, pages 269–298. Springer, 2010.
92. Safdar S, Zafar S, Zafar N, Khan NF. Machine learning based decision support systems (dss) for heart disease diagnosis: a review. Artif Intell Rev. 2018;50(4):597–623.
93. Sarker IH. Context-aware rule learning from smartphone data: survey, challenges and future directions. J Big Data. 2019;6(1):1–25.
94. Sarker IH. A machine learning based robust prediction model for real-life mobile phone data. Internet Things. 2019;5:180–93.
95. Sarker IH. Ai-driven cybersecurity: an overview, security intelligence modeling and research directions. SN Comput Sci. 2021.
96. Sarker IH. Deep cybersecurity: a comprehensive overview from neural network and deep learning perspective. SN Comput Sci. 2021.
97. Sarker IH, Abushark YB, Alsolami F, Khan A. Intrudtree: a machine learning based cyber security intrusion detection model. Symmetry. 2020;12(5):754.
98. Sarker IH, Abushark YB, Khan A. Contextpca: predicting context-aware smartphone apps usage based on machine learning techniques. Symmetry. 2020;12(4):499.
99. Sarker IH, Alqahtani H, Alsolami F, Khan A, Abushark YB, Siddiqui MK. Context pre-modeling: an empirical analysis for classification based user-centric context-aware predictive modeling. J Big Data. 2020;7(1):1–23.
100. Sarker IH, Alan C, Jun H, Khan AI, Abushark YB, Khaled S. Behavdt: a behavioral decision tree learning to build user-centric context-aware predictive model. Mob Netw Appl. 2019; 1–11.
101. Sarker IH, Colman A, Kabir MA, Han J. Phone call log as a context source to modeling individual user behavior. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (Ubicomp): Adjunct, Germany, pages 630–634. ACM, 2016.
102. Sarker IH, Colman A, Kabir MA, Han J. Individualized time-series segmentation for mining mobile phone user behavior. Comput J Oxf Univ UK. 2018;61(3):349–68.
103. Sarker IH, Hoque MM, MdK Uddin, Tawfeeq A. Mobile data science and intelligent apps: concepts, ai-based modeling and research directions. Mob Netw Appl, pages 1–19, 2020.
104. Sarker IH, Kayes ASM. Abc-ruleminer: user behavioral rule-based machine learning method for context-aware intelligent services. J Netw Comput Appl. 2020; page 102762
105. Sarker IH, Kayes ASM, Badsha S, Alqahtani H, Watters P, Ng A. Cybersecurity data science: an overview from machine learning perspective. J Big Data. 2020;7(1):1–29.
106. Sarker IH, Watters P, Kayes ASM. Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. J Big Data. 2019;6(1):1–28.
107. Sarker IH, Salah K. Appspred: predicting context-aware smartphone apps using random forest learning. Internet Things. 2019;8:
108. Scheffer T. Finding association rules that trade support optimally against confidence. Intell Data Anal. 2005;9(4):381–95.
109. Sharma R, Kamble SS, Gunasekaran A, Kumar V, Kumar A. A systematic literature review on machine learning applications for sustainable agriculture supply chain performance. Comput Oper Res. 2020;119:
110. Shengli S, Ling CX. Hybrid cost-sensitive decision tree, knowledge discovery in databases. In: PKDD 2005, Proceedings of 9th European Conference on Principles and Practice of Knowledge Discovery in Databases. Lecture Notes in Computer Science, volume 3721, 2005.
111. Shorten C, Khoshgoftaar TM, Furht B. Deep learning applications for covid-19. J Big Data. 2021;8(1):1–54.
112. Gökhan S, Nevin Y. Data analysis in health and big data: a machine learning medical diagnosis model based on patients' complaints. Commun Stat Theory Methods. 2019;1–10
113. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, et al. Mastering the game of go with deep neural networks and tree search. nature. 2016;529(7587):484–9.
114. Ślusarczyk B. Industry 4.0: Are we ready? Polish J Manag Stud. 17, 2018.
115. Sneath Peter HA. The application of computers to taxonomy. J Gen Microbiol. 1957;17(1).
116. Sorensen T. Method of establishing groups of equal amplitude in plant sociology based on similarity of species. Biol Skr. 1948; 5.
117. Srinivasan V, Moghaddam S, Mukherji A. Mobileminer: mining your frequent patterns on your phone. In: Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing, Seattle, WA, USA, 13-17 September, pp. 389–400. ACM, New York, USA. 2014.
118. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015; pages 1–9.
119. Tavallaee M, Bagheri E, Lu W, Ghorbani AA. A detailed analysis of the kdd cup 99 data set. In. IEEE symposium on computational intelligence for security and defense applications. IEEE. 2009;2009:1–6.
120. Tsagkias M. Tracy HK, Surya K, Vanessa M, de Rijke M. Challenges and research opportunities in ecommerce search and recommendations. In: ACM SIGIR Forum. volume 54. NY, USA: ACM New York; 2021. p. 1–23.

121. Wagstaff K, Cardie C, Rogers S, Schrödl S, et al. Constrained k-means clustering with background knowledge. Icml. 2001;1:577–84.
122. Wang W, Yang J, Muntz R, et al. Sting: a statistical information grid approach to spatial data mining. VLDB. 1997;97:186–95.
123. Wei P, Li Y, Zhang Z, Tao H, Li Z, Liu D. An optimization method for intrusion detection classification model based on deep belief network. IEEE Access. 2019;7:87593–605.
124. Weiss K, Khoshgoftaar TM, Wang DD. A survey of transfer learning. J Big data. 2016;3(1):9.
125. Witten IH, Frank E. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann; 2005.
126. Witten IH, Frank E, Trigg LE, Hall MA, Holmes G, Cunningham SJ. Weka: practical machine learning tools and techniques with java implementations. 1999.
127. Wu C-C, Yen-Liang C, Yi-Hung L, Xiang-Yu Y. Decision tree induction with a constrained number of leaf nodes. Appl Intell. 2016;45(3):673–85.
128. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Philip SY, et al. Top 10 algorithms in data mining. Knowl Inform Syst. 2008;14(1):1–37.
129. Xin Y, Kong L, Liu Z, Chen Y, Li Y, Zhu H, Gao M, Hou H, Wang C. Machine learning and deep learning methods for cybersecurity. IEEE Access. 2018;6:35365–81.
130. Xu D, Yingjie T. A comprehensive survey of clustering algorithms. Ann Data Sci. 2015;2(2):165–93.
131. Zaki MJ. Scalable algorithms for association mining. IEEE Trans Knowl Data Eng. 2000;12(3):372–90.
132. Zanella A, Bui N, Castellani A, Vangelista L, Zorzi M. Internet of things for smart cities. IEEE Internet Things J. 2014;1(1):22–32.
133. Zhao Q, Bhowmick SS. Association rule mining: a survey. Singapore: Nanyang Technological University; 2003.
134. Zheng T, Xie W, Xu L, He X, Zhang Y, You M, Yang G, Chen Y. A machine learning-based framework to identify type 2 diabetes through electronic health records. Int J Med Inform. 2017;97:120–7.
135. Zheng Y, Rajasegarar S, Leckie C. Parking availability prediction for sensor-enabled car parks in smart cities. In: Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), 2015 IEEE Tenth International Conference on. IEEE, 2015; pages 1–6.
136. Zhu H, Cao H, Chen E, Xiong H, Tian J. Exploiting enriched contextual information for mobile app classification. In: Proceedings of the 21st ACM international conference on Information and knowledge management. ACM, 2012; pages 1617–1621
137. Zhu H, Chen E, Xiong H, Kuifei Y, Cao H, Tian J. Mining mobile user preferences for personalized context-aware recommendation. ACM Trans Intell Syst Technol (TIST). 2014;5(4):58.
138. Zikang H, Yong Y, Guofeng Y, Xinyu Z. Sentiment analysis of agricultural product ecommerce review data based on deep learning. In: 2020 International Conference on Internet of Things and Intelligent Applications (ITIA), IEEE, 2020; pages 1–7
139. Zulkernain S, Madiraju P, Ahamed SI. A context aware interruption management system for mobile devices. In: Mobile Wireless Middleware, Operating Systems, and Applications. Springer. 2010; pages 221–234
140. Zulkernain S, Madiraju P, Ahamed S, Stamm K. A mobile intelligent interruption management system. J UCS. 2010;16(15):2060–80.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Quality of Life Research **11**: 763–773, 2002.

Confirmatory factor analysis of the Child Health Questionnaire-Parent Form 50 in a predominantly minority sample

Kimberly A. Hepner & Lee Sechrest
Department of Psychology, University of Arizona, Tucson, USA (E-mail: bissell@u.arizona.edu)

Accepted in revised form 25 April 2002

Abstract

The Child Health Questionnaire-Parent Form 50 (CHQ-PF50; Landgraf JM et al., The CHQ User's Manual. Boston, MA: The Health Institute, New England Medical Centre, 1996) appears to be a useful method of assessing children's health. The CHQ-PF50 is designed to measure general functional status and well-being and is available in several versions to suit the needs of the health researcher. Several publications have reported favorably on the psychometric properties of the CHQ. Landgraf et al. reported the results of an exploratory factor analysis at the scale level that provided evidence for a two-factor structure representing physical and psychosocial dimensions of health. In order to cross-validate and extend these results, a confirmatory factor analysis was conducted with an independent sample of generally healthy, predominantly minority children. Results of the analysis indicate that a two-factor model provides a good fit to the data, confirming previous exploratory analyses with this questionnaire. One additional method factor seems likely because of the substantial similarity of three of the scales, but that does not affect the substantive two-factor interpretation overall.

Key words: Child Health Questionnaire, Confirmatory factor analysis, Quality of life

Abbreviations: BE – behavior; BP – bodily pain; CFA – confirmatory factor analysis; CFI – comparative fix index; CH – change in health item; CHQ-PF50 – child health questionnaire-parent form 50; FA – family activities; FC – family cohesion items; GH – general health scale; MH – mental health; NFI – normed fit index; NNFI – nonnormed fit index; PE – parental impact-emotional; PF – physical functioning; PT – parental impact-time; REB – role/social limitations-emotional/behavioral; RMSEA – root mean squared error of approximation; RP – role/social limitations-physical; SE – self-esteem

Introduction

The Child Health Questionnaire (CHQ) is a recently developed instrument to measure pediatric health outcomes. Several publications document the extensive work to validate the various versions of this questionnaire (e.g., [1–4]). One important consideration in the evaluation of an instrument is construct validity. In the context of assessment, construct validity refers generally to the extent to which the measure assesses the domain, trait, or characteristic of interest [5]. Understanding the latent constructs that influence the observed variables is a critical aspect of construct validity. Construct validity cannot be reduced to a single piece of evidence, but a pattern of evidence can provide support for construct validity. Delineation of the factor structure of an instrument can contribute substantially to the assessment of construct validity. Confirmatory factor analysis (CFA) is particularly useful in that respect. This paper addresses the replicability of the previously obtained factor structure of the CHQ, using different statistical procedures and data obtained from a predominantly minority sample.

The Child Health Questionnaire-Parent Form 50 (CHQ-PF50)

Though the CHQ is relatively new, several studies have reported its use in various versions (e.g., [6–9]). The CHQ is considered a generic (as opposed to disease specific) quality of life measure. In recent years, interest in measuring quality of life as an outcome measure has increased greatly. Though several measures of quality of life have been developed for adults, validated measures available to assess quality of life in children are limited [10]. Although there is some disagreement on just what is meant by quality of life, it has been defined as the 'physical, social and emotional aspects of a patient's well-being that are relevant and important to the individual' [11]. Most definitions are derived from the World Health Organization's definition of health as 'a state of complete physical and mental and social well-being, and not merely the absence of disease or infirmity' [12]. Health-related quality of life has been defined as 'quality of life measures that are likely to be influenced by health interventions' [11]. The CHQ appears to be generally consistent with what is considered a health-related quality of life measure.

The CHQ-PF50 is designed to measure the physical and psychosocial well-being of children 5 years and older. Several forms of the CHQ have been developed that vary in respondent (Parent Form or Child Form) and number of items (PF28, PF50, CF87, PF98). The child form (CHQ-CF87) is appropriate for children 10 years old or older. The CHQ-PF50 was empirically derived from the full-length CHQ-PF98 in 1994. The CHQ-PF50 assesses several areas including general health, change in health, physical functioning, bodily pain, limitations in school work and activities with friends, behavior, mental health, self-esteem, time and emotional impact on the parent, limitations in family activities and family cohesion. The health concepts measured by the CHQ-PF50, including the number of items contributing to each concept, are further described in Table 1.

A four-week recall period is used for all scales except for the change in health (CH) item, the family cohesion (FC) items, and the general health (GH) scale. Scoring provides a score for each scale. In addition, two summary scores are provided: a physical health summary score and a psychosocial health summary score. The CHQ has been translated into several languages including American–Spanish, Canadian–French, Finnish, French, German, Dutch, Italian, Greek, Honduran, Mexican, Norwegian, Portuguese, and Swedish [10]. A more extensive description of development of the CHQ, the health concepts assessed, and interpretation of scale scores is given in the manual [1].

Previous factor analyses of the CHQ

Development and validation of the CHQ included factor analyses of the correlations among the scales. Exploratory factor analytic techniques, specifically principal components analyses, provided evidence for a two-factor structure representing physical and psychosocial functioning. These results were based on a sample of 914 children from both general population and specific condition (asthma, ADHD, cystic fibrosis, epilepsy, rheumatoid arthritis, psychiatric problems) groups. Four scales (physical functioning, role/social-physical, general health perceptions, bodily pain) loaded strongest on the physical factor. Four scales (role/social-emotional/behavioral, self-esteem, mental health, behavior) loaded strongest on the psychosocial factor. Two scales (parental impact-time, parental impact-emotional) loaded on both factors, but showed stronger loadings on the psychosocial health factor.

Replication of factor analysis

Attempts to replicate results from previous factor analyses are important for several reasons. Gorsuch [13] emphasized the importance of assessing how well factors can be replicated and how invariant the factors are across samples. Replication addresses how well factors generalize across samples drawn from the same population. Invariance, in contrast, addresses how well factors generalize across the specific variables or different samples. The concern of this paper is chiefly the replication of the factor structure obtained previously by other researchers.

Table 1. Health concepts measured with number of items in the CHQ-PF50

Health concept	Number of items	Brief description
Physical functioning (PF)	6	Measures the presence and extent of physical limitations due to health related problems
Role/social limitations-physical (RP)	2	Measures limitations in the kind, amount and performance of school work and activities with friends due to physical health problems
General health perceptions (GH)	6	Measures perceptions concerning the child's overall health in the past, present, and future
Bodily pain/discomfort (BP)	2	Measures the intensity and frequency of general pain or discomfort
Parental impact-time (PT)	3	Measures limitations in personal time experienced by the parent/guardian due to child's physical health, emotional well being/general behavior, and attention or learning abilities
Parental impact-emotional (PE)	3	Measures the amount of distress experienced by the parent/guardian related to the child's physical health, emotional well being/general behavior, and attention or learning abilities
Role/social limitations-emotional/behavioral (REB)	3	Measures limitations in the kind, amount and performance of school work and activities with friends due to emotional or behavioral difficulties
Self-esteem (SE)	6	Measures several dimensions of self-esteem including satisfaction with school and athletic ability, looks/appearance, ability to get along with others and family, and life overall
Mental health (MH)	5	Measures the frequency of both positive and negative states including anxiety, depression, and positive states
General behavior (BE)	6	Measures overt behavior as a component of mental health including behavior problems and ability to get along with others
Family activities (FA)	6	Measures the frequency of disruption in 'usual' family activities due to the child's health or behavior
Family cohesion (FC)	1	Measures the family's ability to get along
Change in health (CH)	1	Subjective assessment of child's health as compared to one year ago

(Modified with permission from the CHQ User Manual [1], pp. 33–38.)

Several circumstances influence the results of factor analysis, including sample size, the communalities among the variables, the number of variables per factor, and the factor analytic method used. In factor analysis, sample size is usually discussed in terms of the number of cases per variable. Gorsuch [13] suggested that the absolute minimum ratio is five individuals per variable and no fewer than 100 individuals, although there is very little evidence supporting any one notion [14].

Communalities also influence the results, and subsequent replicability, of factor analyses since communalities reflect the strength of the phenomena and the accuracy of measurement [13]. Factor loadings become more replicable as communalities increase. The likelihood of replication also increases as the number of variables per factor increases. Gorsuch [13] suggested a minimum of four variables per factor, but acknowledged that some CFAs may prove to be exceptions to this rule.

Confirmatory factor analysis

CFA requires the specification of a factor model, including the number of factors and the pattern of zero and nonzero loadings on those factors. A small number of theory-driven competing models might be specified as well. CFA provides information on how well the hypothesized model explains the relations among the variables. CFA has the advantages of allowing hypothesis testing on the data and may offer fewer opportunities to capitalize on chance because of *a priori* model specification [15]. The extensive prior work to develop the theory underlying the CHQ and previous factor analytic studies provide a sound basis on which to test replicability of factor structure by CFA.

The question posed in the research reported here was more demanding than simply whether the factor structure for the CHQ would be replicable. In addition we were interested in using the CHQ with a sample of children from a low income, predominantly minority, generally healthy population, so the test of replicability was moderately stringent.

Method

Participants

Participants were selected from community health centers in Tucson, Arizona. Two methods of participant solicitation were used. One sample was collected via mail and one sample was collected in person. These two samples were later combined to create the sample used for the CFA. As part of a pilot evaluation of a health insurance program for children from low-income families, questionnaires were mailed to parents of children enrolled in the program. To maintain confidentiality, prepared participation materials were provided to a community health center that then addressed and mailed them. Parents with children of any age received a questionnaire; however, the CHQ-PF50 is considered appropriate only for children at least 5 years old. The response rate for this sample was 51% (100 mailed, 51 completed). Responses for children younger than five (6) and questionnaires with missing data on the scales (2) were eliminated from the sample, leaving a sample size of 43.

A second sample consisted of participants approached in the pediatrics office of a community health center. Parents were invited to participate if they were the parent or legal guardian of a child who was 5 years or older. The response rate for this sample was 89.4% (132 approached, 118 completed). Ten questionnaires were eliminated because of missing data on one or more of the scales, leaving a sample size of 108. Participants were treated in accordance with the 'ethical principles of psychologists and code of conduct' [16].

The two samples are highly similar in terms of sociodemographic variables, and both contacts were in the context of health care. There is no reason to suppose that either the method of contacting parents nor the circumstances under which they filled out the questionnaire would have affected the factor structure of the instrument, the principle focus of this study. The analyses were, thus, based on the combined sample of 151 participants (see Table 2). The total sample is heavily weighted toward female (mother) respondents, lower educational levels, and Hispanic ethnicity. Table 2 also provides physical health and psychosocial health summary scores. The samples were compared using a series of χ^2 tests (for categorical demographic characteristics) and ANOVAs (for summary scores). These tests indicated only one significant difference between the two samples: sample two included more respondents from a minority background [χ^2 (1, $n = 151) = 22.23, p < 0.01$], primarily due to a larger number of respondents reporting Hispanic ethnicity.

Design

Though two methods of data collection were used, all administrations of the CHQ-PF50 followed the recommendations provided by the manual [1]. Participants who responded via mail received study materials including a cover letter explaining the project, a $5 incentive, an informed consent form, a stamped, addressed envelope, and a questionnaire. The questionnaire included the CHQ-PF50, as well as additional questions designed to assess the impact of health insurance on family health and dynamics. The CHQ-PF50 was reproduced exactly including instructions, question order, and headings. Participants received an additional $10 for returning a completed survey directly to the investigator.

Table 2. Sociodemographic characteristics of the samples

	Sample one	Sample two	Combined sample
Female respondent (mother)	92.9%	84.3%	86.7%
Biological parent	97.6%	89.7%	91.9%
High school education or less	61.0%	58.8%	59.5%
Married	40.5%	49.5%	47.0%
Minority racial background	51.2%	87.0%	76.8%
Hispanic racial background	44.2%	65.7%	59.6%
Physical health summary score	49.0	45.5	46.5
Psychosocial health summary score	51.0	48.8	49.4

Participants who responded in-person were approached in the waiting room prior to seeing a doctor. Participant materials include an informed consent form and a questionnaire, the CHQ-PF50. Again, the CHQ-PF50 was reproduced exactly, including instructions, question order, and headings. The cover letter was replaced by an oral explanation of the project by the investigator. Participants were entered in a drawing for $150 for returning a completed survey.

Introduction to the questionnaire, adapted from that provided in the CHQ manual [1], was similar whether it was provided in written (via cover letter) or verbal form. The general purpose of the study was described. Participants were reminded to read the instructions and that there were no right or wrong answers. Participants were also asked not to share their responses or ask for help from any of their family members. The investigators were available by telephone or in person to answer questions, depending on the method, with only occasional questions during the in-person method.

Scoring

Completed questionnaires were scored according to the SAS [17] protocol provided in the CHQ Manual [1], which includes a detailed, item-by-item description of the scoring method. The user is instructed to recheck questions, item stems, response choices, and response values to ensure they are taken verbatim from the manual. The data were checked for any out-of-range values; none were found. Items were recoded to ensure that for all items a higher score indicated better health. Items were also recalibrated to account for differing response continua. Scale scores are only calculated for those individuals for whom half or more of the items in the scale had been answered. Raw scale scores were calculated by computing the algebraic mean of the completed items. The raw scores were then transformed so that the scale scores range from 0 to 100. Eleven scale scores are provided (PF, REB, RP, BP, BE, MH, SE, GH, PE, PT, and FA). Two summary scores were also calculated: a physical summary score and a psychosocial summary score. These summary scores are calculated by standardizing the scales based on general population and clinical samples, aggregating the scales using factor weights from these samples, and transforming the scores to have a mean of 50 and a standard deviation of 10.

Previous analyses did not include the family activities (FA) scale nor the FC and the CH items because these items were not included in some clinical samples. For purposes of replication, these items were omitted from this analysis as well. Therefore, the analysis was based on the 10 remaining scales (PF, RP, GH, BP, PT, PE, REB, SE, MH, BE). The two summary scores are calculated using only these 10 scales.

Statistical analyses

Factor analyses are facilitated if conducted with standardized variables [18]; therefore, scale scores were standardized and *z*-scores served as variables in the CFA. CFAs were performed using EQS [19], a causal modeling program.

The primary task in testing confirmatory factor analytic models is to determine the goodness of fit between the hypothesized model and the sample data [20]. The adequacy of model fit was evaluated using the χ^2 statistic, the Bentler-Bonnet normed fit index (NFI), the nonnormed fit index (NNFI), the comparative fit index (CFI), and Steiger's root mean square error of approximation (RMSEA). χ^2 reflects the statistical goodness of fit of the observed matrix compared to the expected matrix predicted by the hypothesized model. A significant χ^2 value suggests that the hypothesized factor model is not adequate, but the χ^2 statistic is sensitive to sample size. With large samples, trivial discrepancies can lead to rejection of an otherwise good model; with small samples, χ^2 can be non-significant even when the model does not fit well [18]. Because of the sensitivity of the χ^2 statistic, it is important to use some 'practical' indices of fit to supplement evaluation of the proposed model. Both the NFI and the CFI range from 0 to 1.00, with a value greater than 0.90 being generally taken to indicate an acceptable fit to the data [21]. These two fit indices are based on a comparison of the hypothesized model with the null model (i.e., all correlations among the variables are 0). The NFI has been shown to underestimate fit in small samples. The CFI, on the other hand, was designed the take sample size into account. Therefore, the CFI should be the primary index used when evaluating model fit [20]. An additional index that should be taken into account is RMSEA. Whereas previous indices discussed are considered goodness-of-fit indices, RMSEA is a 'badness of fit' index [18] because a value of 0 indicates perfect fit. This index is a population-based index and therefore, relatively insensitive to sample size. RMSEA values below 0.10 may be considered good, and the lower the better [18].

Results

The primary goal of the current research was replication of the factor structure, but an advantage of CFA is that it makes possible the comparison of competing models. Several different models were tested and are reported on here.

Model A: two factors

Because replication of the previously obtained factor structure was the primary interest in the current study, a two-factor model was tested initially (Figure 1). The two factors represent physical health and psychosocial health. Four variables were hypothesized to load on the first factor: physical functioning (PF), role functioning-physical (RP), general health (GH), and bodily pain (BP), with the other six variables hypothesized to load on the second factor: parental impact-time (PT), parental impact-emotional (PE), role functioning-emotional/behavioral (REB), self-esteem (SE), mental health (MH), and behavior (BE). The CFA specifies that each scale should have a nonzero loading on its hypothesized factor and a 0 loading on the other factor. It was also expected that the two factors would be correlated. Finally, the initial model assumed that the measurement error terms would be uncorrelated.

Model A did not provide a particularly good fit to the data (Table 3). The CFI of 0.767 indicates that the hypothesized model is not an adequate representation of the observed data. The significant χ^2 value also indicates some misfit ($\chi^2 = 166.61$, df $= 34, p < 0.01$). Other indices indicated poor fit as well (NFI = 0.728, NNFI = 0.691, RMSEA= 0.161). Before attempting to improve the model fit, it was important to rule out other plausible models. These two alternative models will be described before returning to improving the fit of Model A.

Alternative models

The poor fit of the two-factor model, along with the correlations between the two factors suggests the possibility that the data might be explained by a single factor model, perhaps representing global health. Thus, it was hypothesized that each scale would have a nonzero loading on the factor. In addition, the model specified that the measurement error terms would be uncorrelated.

The single factor model (Model B) proved a poor fit to the data with a CFI of 0.679. The χ^2 value also indicated poor fit ($\chi^2 = 217.49$, df $= 35$, $p < 0.01$). Other fit indices indicated poor fit as well (NFI = 0.645, NNFI = 0.587, RMSEA = 0.187). When models are nested, as are the models in this series of CFAs, the χ^2 statistic can be used to assess the difference in χ^2 between two nested models. When assessing the differences between models, a significant χ^2 indicates that one of the models represents a better fit than the other. The results shown in Table 4 suggest that the two-factor model (Model A) is probably a better fit to the data than a single factor model (χ^2 difference $= 50.88$, df $= 1$, $p < 0.01$).

Previous results, as discussed above, indicated that the two parental impact scales, though loading most strongly on the psychosocial factor, also had notable secondary loadings on the physical factor. Therefore, it was desirable to test a two-factor model with cross-loadings (Model C) that permitted these two scales to have nonzero loadings on each of the two factors. For the other eight scales, it was hypothesized that each scale would have a nonzero loading on its hypothesized factor and a zero loading on the other factor. It was also predicted that the two factors would be correlated. Finally, the model indicated that the measurement error terms would be uncorrelated. An initial run of this model indicated that the model did not provide a good fit to the data. The two-factor model with cross loadings (Model C) had a CFI of 0.767 and a significant χ^2 value ($\chi^2 = 164.50$, df $= 32$, $p < 0.001$). Other indices indicated poor fit as well (NFI = 0.732, NNFI = 0.672, RMSEA = 0.166). The fit indices are similar to those of Model A (χ^2 difference $= 2.11$,

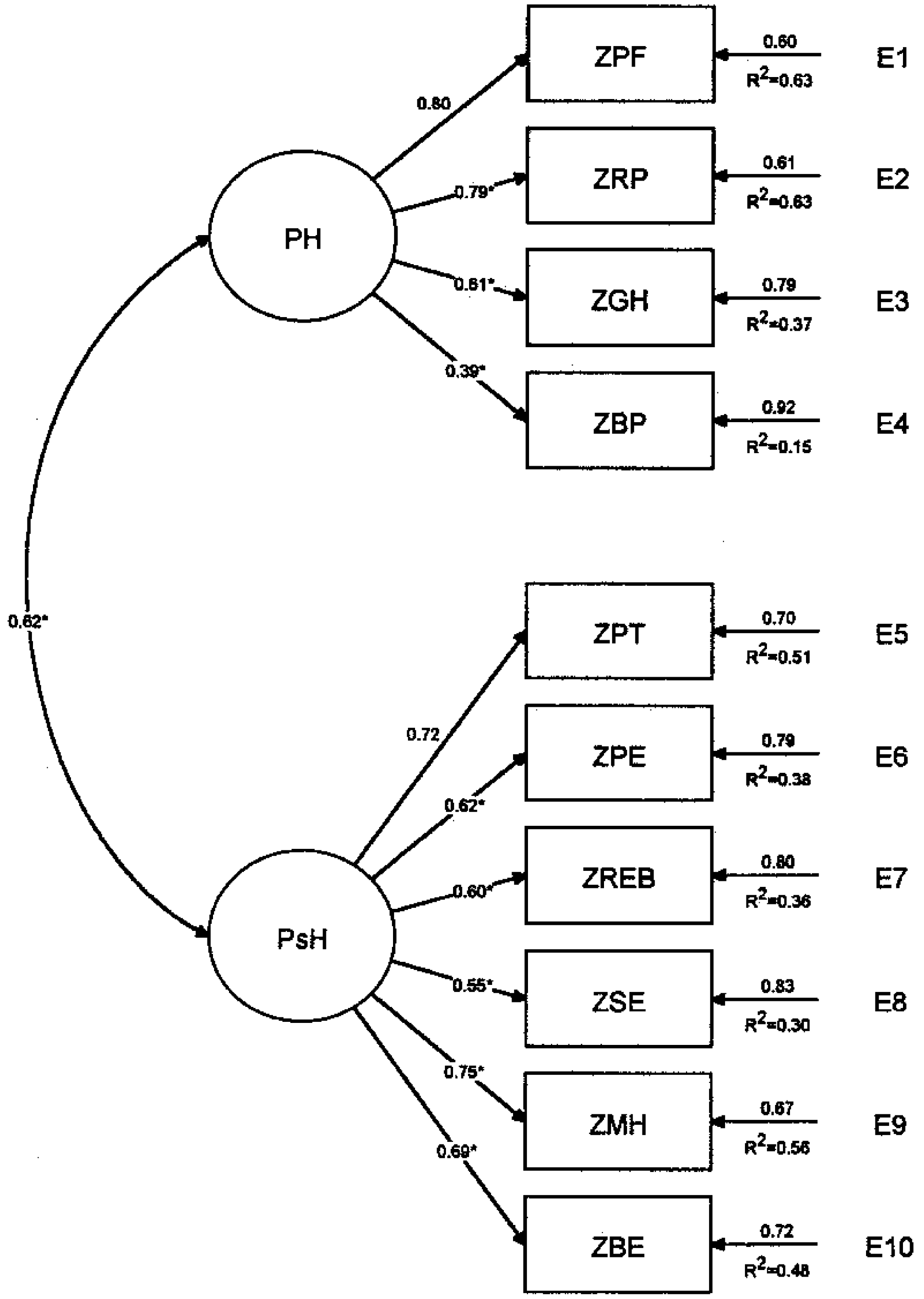


Figure 1. Model A: a two-factor model.

df = 2, $p > 0.05$); however, Model A was favored over Model C as a more parsimonious explanation of the observed data.

Though Model A represented a better and more parsimonious explanation of the data than Model B or C, it still did not provide a satisfactory fit. Therefore, it was desirable to modify the primary hypothesized model in order to improve the model fit. Inspection of the standardized residuals led to a closer examination three scales: PF, REB, and RP. A portion of the variance in the data involving these three variables was not accounted for by the

Table 3. Overall goodness-of-fit indices for the CHQ

Codes	Fit indices						
	χ^2	df	p (Ho)	CFI	NFI	NNFI	RMSEA
A	166.61	34	0.00	0.767	0.728	0.691	0.161
B	217.49	35	0.00	0.679	0.645	0.587	0.187
C	164.50	32	0.00	0.767	0.732	0.672	0.166
D	58.71	31	0.00	0.951	0.904	0.929	0.077
E	58.71	31	0.00	0.951	0.904	0.929	0.077

Model A – two-factor model, Model B – single-factor model, Model C – two-factor with cross-loadings model, Model D – two-factor with correlated errors model, Model E – three-factor model.

Table 4. Model comparisons

Codes	χ^2 difference	df	p (Ho)
A vs. B	50.88	1	<0.01**
A vs. C	2.11	2	>0.05
A vs. D	107.90	3	<0.01**
A vs. E	107.90	3	<0.01**

hypothesized model, but it did not make theoretical sense to allow these variables to cross-load on the other factor.

Upon closer inspection of these three scales, it became apparent that these scales might be correlated for methodological reasons. These three scales appear together at the beginning of the questionnaire, just after a single overall health question. Portions of these scales are reproduced in Table 5 in the order in which they appear in the questionnaire. In addition, the structures of the questions, particularly the answer choices, are very similar. Finally, and perhaps most importantly, the scales all address limitations that were experienced by children due to different aspects of their health.

Therefore, Model A was altered to create Model D (Figure 2). Model D is identical to Model A, with the exception that the measurement error terms for PF, REB, and RP were allowed to covary. Model D had a significantly better fit than Model A (χ^2 difference = 107.90, df = 3, $p < 0.01$). The CFI for Model D was 0.95, indicating that this model provides a good fit to the data. The χ^2 value for this model is still significant ($\chi^2 = 58.71$, df = 31, $p < 0.01$), suggesting that there still may be some misfit. Other indices indicated a good fit (NFI = 0.904, NNFI = 0.929, RMSEA = 0.077).

Though Model D appeared to be the model of choice, one final model was tested to examine the

Table 5. Sample items from three scales

Physical functioning (PF)	During the past 4 weeks, has your child been limited in any of the following activities due to health problems? Doing things that take a lot of energy, such as playing soccer or running?*
Role/social limitations-emotional/behavioral (REB)	During the past 4 weeks, has your child's school work or activities with friends been limited in any of the following ways due to EMOTIONAL difficulties or problems with his/her BEHAVIOR? Limited in the KIND of schoolwork or activities with friends he/she could do*
Role/social limitations-physical (RP)	During the past 4 weeks, has your child's school work or activities with friends been limited in any of the following ways due to problems with his/her PHYSICAL health? Limited in the KIND of schoolwork or activities with friends he/she could do*

(From the CHQ-PF50 in the CHQ User Manual [1], pp. 364–365, with permission.)

* These scales share a common response set that ranges from 'Yes, limited a lot' to 'No, not limited'.

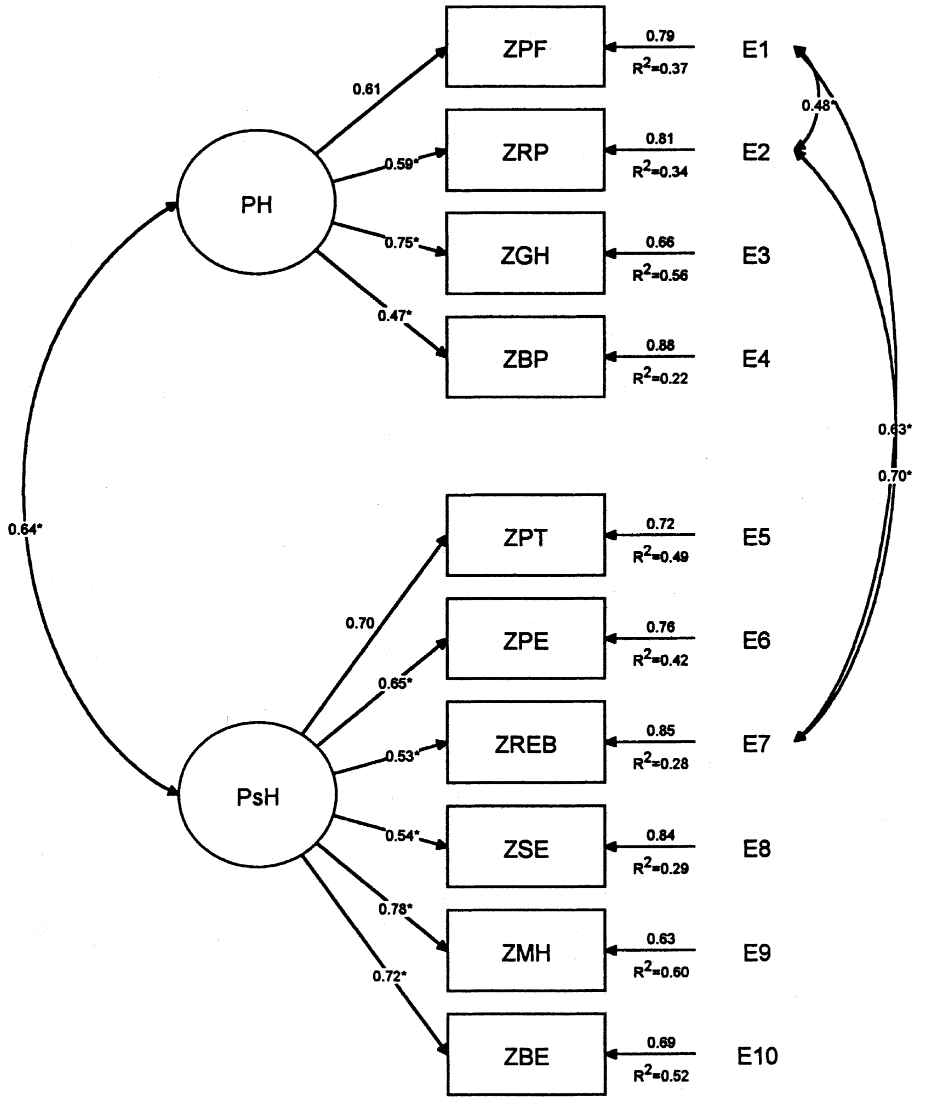


Figure 2. Model D: a two-factor with correlate errors model.

possibility of a third factor that would represent a method factor. This model, Model E, was identical to Model A with the exception that the three scales that appeared to correlate for methodological reasons were allowed to cross-load on a third factor. When Model E was tested, it yielded fit indices identical to those of Model D.

Discussion

It appears from the analyses that both Models D and E provide an adequate representation of the data according to the indicators of fit used to evaluate the models. Consideration of the theoretical bases for these two models, however, may result in favoring Model D over E. By specifying a third factor that appears to represent a method factor, Model E gives too much credence to the presence of presumably unwanted method variance. In theory, the presence of correlated error terms should represent the case in which unwanted variance intrudes on the variance that we are interested in (i.e., wanted variance). There is no reason to believe that the variance associated with methodological factors (i.e., question similarity and question proximity) is desired or is important enough to include as a factor in the measurement model of the CHQ. Rather, the measurement model initially hypothesized by the developers of the questionnaire remains intact with the addition of three correlated error terms.

The question remains, then, what can we learn from Model D about the measurement model of the CHQ. First, the factor structure obtained in previous analyses has been replicated. This finding supports the proposition that the CHQ does indeed assess two constructs that appear to represent physical health and psychosocial health. It is, we think, important that the factor structure was replicated in a sample of relatively healthy children, a large proportion of whom were from minority backgrounds.

This paper is not a test of factorial invariance because the analysis was not conducted on a systematically selected portion of the sample (i.e., Hispanics only). The current analysis, however, may provide an interesting clue as to what the results of a test of factorial invariance may reveal. Of the sample used in the current study, approximately 77% reported a minority ethnic background. The results of the present study suggest that the factor structure would be invariant in at least one minority population.

The second important finding is that method variance may need to be considered in the measurement model of the CHQ. The primary investigator in the development of the CHQ indicated that in developing the questionnaire these similar scales were intentionally placed together at the beginning (J.M. Landgraf, personal communication). Because these scales had an obviously similar structure they were placed together at the beginning of the questionnaire to make it easier for the person completing the CHQ to get started quickly. Clearly, the ease of beginning is a potential benefit of the choice to group them. There are some drawbacks to that decision as well. Placing similar scales together increases the likelihood that the scales will covary to some degree because of ‘method’ or (unwanted) variance. Stated another way, the answers chosen on these scales may be influenced unduly by the similarity in the question method rather than the similarity of the question content. For a more extensive discussion of method variance (see Ref. [21]).

It is only at a conceptual level that the difference between Models D and E is of any consequence. Whether one regards the residual correlations between the three specific items as shared error or as the result of a separate method factor does not affect any further use of those variables, e.g., in deciding whether and how to ‘correct’ for the unwanted correlations among them. The magnitude of the unwanted component is of some importance; if it is small, the decision whether to try to correct for it is inconsequential. If, as in the present case, the correlated error, or method variance, is substantial (as suggested by the values of the path coefficients), then partialing that term out of the overall factors seems warranted.

The CFA results discussed here represent but a single set of findings that are meaningful *only* in the context of other validation work on this questionnaire. Although conceptually Models D and E are distinct but lacking in differential analytic consequences, the size of the error or method effect that they represent is not trivial, and cross-validation of the values for the path coefficients estimated here should be attempted. If corrections for error or methods effects are to be done, it is critical that the corrections be of reasonable accuracy. The analyses presented in this paper are only a modest, although we think important, step in the continuing validation of the CHQ. In the meantime, the CHQ appears to be a useful tool for assessing the health of children.

Acknowledgement

This work was supported in part through a grant from the Flynn Foundation.

References

1. Landgraf JM, Abetz L, Ware JE. The CHQ User’s Manual. 1st ed., Boston, MA: The Health Institute, New England Medical Center, 1996.
2. Asmussen L, Olson LM, Grant E, Fagan J, Landgraf JM, Weiss K. Test–retest reliability of the Child Health Questionnaire (CHQ) in a sample of moderate and low-income, urban children with asthma. J Allergy Clin Immunol 1999; 103(Suppl. Part 2)(1): 654.
3. Landgraf JM, Abetz LN. Functional status and well-being of children representing three cultural groups: Initial self-

reports using the CHQ-CF87. Psychol Health 1997; 12(6): 839–854.

4. Landgraf JM, Maunsell E, Speechley KN, et al. Canadian–French, German and UK versions of the Child Health Questionnaire: Methodology and preliminary item scaling results. Qual Life Res 1998; 7(5): 433–445.
5. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychol Bull 1955; 52: 281–302.
6. Gilliam F, Wyllie E, Kashden J, et al. Epilepsy surgery outcome: Comprehensive assessment in children. Neurology 1997; 48(5): 1368–1374.
7. Levi RB, Drotar D. Health-related quality of life in childhood cancer: Discrepancy in parent-child reports. Int J Cancer 1999; 12(Suppl.): 58–64.
8. Sawyer M, Antoniou G, Toogood I, Rice M. A comparison of parent and adolescent reports describing the health-related quality of life of adolescents treated for cancer. Int J Cancer 1999; 12(Suppl.): 39–45.
9. Stewart MG, Friedman EM, Sulek M, et al. Quality of life and health status in pediatric tonsil and adenoid disease. Arch Otolaryngol Head Neck Surg 2000; 126(1): 45–48.
10. Connolly MA, Johnson JA. Measuring quality of life in paediatric patients. Pharmacoeconomics 1999; 16(6): 605–625.
11. Glossary. Pharmacoeconomics 1998; 13: 109–110.
12. World Health Organization. World Health Organization Constitution. Basic Documents. Geneva: World Health Organization, 1948.
13. Gorsuch RL. Factor Analysis. (2nd ed.) Hillsdale, NJ: Lawrence Erlbaum, 1983.
14. MacCallum RC, Widaman KF, Zhang S, Hong S. Sample size in factor analysis. Psychol Meth 1999; 4(1): 84–99.
15. Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ. Evaluating the use of exploratory factor analysis in psychological research. Psychol Meth 1999; 4(3): 272–299.
16. American Psychological Association. Ethical principles of psychologists and code of conduct. Amer Psychol 1992; 47: 1597–1611.
17. SAS Institute Inc. SAS/STAT User's Guide, Version 6, 4th ed. Vol. 1. Cary, NC: SAS Institute Inc, 1989.
18. Loehlin JC. Latent Variable Models: Factor, Path, and Structural Analysis. Mahwah, NJ: Lawrence Erlbaum, 1998.
19. Bentler PM. EQS Structural Equations Program Manual. Los Angeles: BMDP Statistical Software, 1989.
20. Byrne BM. Structural Equation Modeling with EQS and EQS/Windows: Basic Concepts, Applications, and Programming. Thousand Oaks, CA: SAGE, 1994.
21. Sechrest L, Davis MF, Stickle TR, McKnight PE. Understanding 'method' variance. In: Leonard Bickman (ed.), Research Design: Donald Campbell's Legacy. Thousand Oaks, CA: Sage, 2000; 63–87.

Address for correspondence: Kimberly Bissell, Department of Psychology, University of Arizona, Tucson, AZ 85721, USA
Phone: + 1-520-621-5463; Fax: + 1-520-621-6320
E-mail: bissell@u.arizona.edu

Journal of Clinical Epidemiology 62 (2009) 1138−1147

METHODOLOGICAL INNOVATIONS

Classification and regression tree uncovered hierarchy of psychosocial determinants underlying quality-of-life response shift in HIV/AIDS

Yuelin Li[a,*], Bruce Rapkin[b]

[a]Department of Psychiatry and Behavioral Sciences, Memorial Sloan-Kettering Cancer Center, 641 Lexington Avenue, 7th Floor, New York, NY 10022, USA
[b]Department of Epidemiology and Population Health, Albert Einstein College of Medicine (AECOM) of Yeshiva University, Bronx, NY 10461, USA

Accepted 31 March 2009

Abstract

Objectives: Rapkin and Schwartz define response shift as otherwise unexplained, discrepant change in health-related quality of life (HRQOL) that is associated with change in cognitive appraisal. In this article, we demonstrate how a recursive partitioning (rpart) regression tree analytic approach may be used to explore cognitive changes to gain additional insight into response-shift phenomena.

Study Design and Setting: Data are from the ''Choices in Care Study,'' an evaluation of HIV+ Medicaid recipients' experiences and outcomes in care ($N = 394$). Cognitive assessment was based on the QOL appraisal battery. HRQOL was measured by the SF-36 Health Survey, version 2 (SF-36v2).

Results: We used rpart to examine 6-month change in SF-36v2 mental composite score as a function of changes in appraisal, after controlling for patient characteristics, health changes, and intervening events. Rpart identified nine distinct patterns of cognitive change, including three associated with negative discrepancies, four with positive discrepancies, and two with no discrepancies.

Conclusion: Rpart classification provides a nuanced treatment of response shift. This methodology has implications for evaluating programs, guiding decisions, and targeting care.

Keywords: Response shift; Health-related quality of life; Classification and regression trees; Segmentation strategies; Idiographic quality of life assessment; Rpart

1. Introduction

Converging evidence shows that response shift can strongly affect how individuals appraise their health-related quality of life (HRQOL) [1−4]. Response shift typically appears in counterintuitive findings—individuals with severe chronic illnesses reporting equal or better HRQOL scores than healthy individuals or individuals with less severe illness (e.g., [5,6]). For example, the general public assigned a 0.39 HRQOL to dialysis, whereas dialysis patients assigned their own HRQOL at 0.56 (on a 0−1 scale where 0 represents death and 1 represents perfect health) [7]. This and similar paradoxical findings bring into question what HRQOL assessments are really measuring. Measurement imprecision and response bias do not fully explain the phenomena [1]. The theory of response shift posits that it constitutes a change in the meaning of one's self-evaluation of the QOL construct because of recalibration, repriortization, and/or reconceptualization [4,8]. These constructs are related to work on idiographic QOL assessment [2,9−11]. The theoretical and measurement foundations of these constructs are well documented [3,4,12].

Rapkin and Schwartz [1] describe the assessment strategies to probe respondents on their evaluation of the meaning of QOL. They propose operationalizing response shift as change in HRQOL that cannot be explained by changes in overt health status, resources or life events, but that can be associated with change in cognitive appraisal. Their data-analytic strategy was based primarily on linear regression to estimate the extent to which residual QOL changes are associated with appraisal change [2]. However, there is no intrinsic reason that these relationships must be linear. Rapkin and Schwartz' notion of a final ''combinatorial algorithm'' that people use to summarize their experiences into HRQOL ratings explicitly posits complex interactions among constituents of appraisal. For example, an individual may report better QOL than expected given their health status by ignoring problems, emphasizing positive experiences, selecting favorable targets for self-comparison, and/or focusing on less ambitious goals. Each of these processes may operate alone or in combination to represent distinct types of response shift.

There are obvious drawbacks to using linear regression to examine relationships involving appraisal processes that are

[*] Corresponding author. Tel: +1-646-888-0047; fax: +1-212-888-2959.

E-mail address: liy12@mskcc.org (Y. Li).

0895-4356/09/$ − see front matter © 2009 Elsevier Inc. All rights reserved.
doi: 10.1016/j.jclinepi.2009.03.021

What is new?

We used the rpart classification and regression tree to uncover the hierarchy of cognitive determinants underlying QOL response shift in HIV/AIDS. Highest in the hierarchy was the reduction of the salience of negative experiences between baseline and 6 months follow-up. A moderately large reduction (e.g., avoid thinking about things that are disappointing, worrisome, or difficult) was associated with a positive response shift (i.e., better QOL than expected by overt health status) in overall mental health. A combination of other cognitive variables also came into play. For example, increased concerns about monetary obligations and other external demands were associated with a negative response shift. These findings may help meet patients' needs, perhaps by linking patients with resources and support. Change in the content and process of cognitive appraisal is a worthy patient-reported outcome domain in its own right. The rpart classification technique provides a nuanced interpretation of response shift.

intrinsically nonlinear. Classification and regression trees (CART) methods are a suitable alternative to linear regression in elucidating potentially complex interactions [13]. Using an iterative algorithm, respondents are classified into increasingly homogeneous subgroups with similar changes in cognitive appraisal profiles, allowing a more nuanced interpretation of how cognitive appraisal can influence HRQOL. The broad goal of this article is to demonstrate an empirical technique to identify prevalent patterns of cognitive changes that can account for residual variance in HRQOL change scores. Patterns of appraisal identified in this way represent different manifestations of response shift.

2. Methods

2.1. HIV/AIDS choices in care study

The study was developed by investigators at our respective institutions in conjunction with the New York State Department of Health AIDS Institute to evaluate the impact of the HIV Special Needs Plans, as part of an evaluation of patient-reported outcomes and experiences in care reported by HIV+ Medicaid recipients in New York State. Detailed data collection plans are summarized elsewhere [14,15]. Institutional Review Boards approved the study.

Interviews were conducted in either Spanish or English, in person or by telephone, according to patient preference. The primary HRQOL assessment was the 36-Item Short Form Health Survey, version 2 (SF-36v2) [16], assessed at baseline (approximately 6 weeks post enrollment), and at 6 and 12 months post baseline. Changes in cognitive appraisal processes were assessed at these time points using the QOL appraisal battery [1]. The baseline interview also included measures of demographics, behavioral risks, and health history.

2.2. QOL appraisal battery

After Rapkin and Schwartz [1], the QOL appraisal battery included four components: (1) persons' *frame of reference* for considering HRQOL as assessed by six probes designed to tap different motivational themes, including achievement, maintenance, prevention, problem solving, disengagement, and acceptance. For example, respondents were asked about ''the main things you want to accomplish,'' ''problems you want to solve,'' and ''things you are trying learn to accept,'' to have their best possible QOL. Verbatim responses to these probes, or ''goal statements,'' were coded and analyzed to extract ''goal attributes'' (described below). Additionally, we assessed (2) how persons *sample experiences* within that frame, assessed by 13 items on, for example, whether or not the persons evaluated HRQOL by ''thinking about the worst possible moments'' within that frame, (3) how persons evaluate experiences using different *standards of comparison* by nine items on, for example, whether they compared themselves with ''other people living with HIV,'' and (4) how persons summarize and combine evaluations to describe HRQOL by using a combinatory algorithm of 16 items on, for example, whether they were thinking about ''how well you've been doing, how hard it has been, both or neither?''.

2.3. Coding and summarizing goal statements to assess frame of reference

From the open-ended assessments of frame of reference, we collected over 6,700 goal statements at baseline and 6 months (plus an additional 1,458 from our first wave of 12-month follow-up interviews, which were coded in this group, but not reported here). Content analysis of these responses was accomplished through a two-stage process that is briefly summarized below. Complete documentation of goal coding, *kappa* reliability, and components analysis are available from the authors on request.

In the first step, we selected at random just over one of three of all responses (2,638). Each selected goal statement was given to two of 13 judges (students and faculty in our department), after an allocation scheme to ensure that an equal number of overlapping goals were assigned to each pair of judges. Each judge independently sorted about 405 goal statements into homogeneous categories, with the sole criterion being that statements within a category must be ''similar in all important ways.'' Judges then recorded the ''goal attributes'' that they used to make distinctions among categories, including life domains, motivations, and health relevance. After completion of independent sorting, all judges met to compare their derived dimensions. In general, there was strong agreement in the major distinctions among life domains and in prevalent fine-grain distinctions. Judges primarily differed in how specific to be in certain subdomains (e.g., to distinguish concerns about specific family members from those pertaining to the family in general). Based on this discussion, we derived a consensus set of 24 binary goal attributes. All goal statements could be characterized using combinations of these goal attributes. We calculated *kappa* for each of the 24 codes, to determine whether or not pairs of judges agreed on the presence or absence of each goal attribute in their initial sort of goal statements. Collapsing across dyads, we found that 11 of 24 categories exceed *kappa* = 0.70, another four exceed *kappa* = 0.59, six exceeded *kappa* = 0.35, and three codes (representing only 3.76% of coded statements) did not differ from chance.

After derivation of goal attributes, the remaining 2 of 3 (5,148) goal statements were assigned to 11 judges. We assigned a random 20% of these goal statements (1,030) as a reliability sample, allocated evenly to all possible pairs of judges. Reliability coefficients for 13 of 24 categories exceeded *kappa* = 0.70, another six exceed *kappa* = 0.50, three exceeded an acceptable level of *kappa* = 0.39, and two categories (representing only 1.19% of coded statements) were not different from chance. Based on these results, final goal attributes were coded for each goal statement. Note that in final coding, we resolved disagreements among judges by assuming that differences were resulting from errors of omission (one judge indicated a code that the other did not).

Our next step involved combining scores across all of individual's goal statements at baseline and separately at 6-months to characterize current priorities and concerns at each time of measurement. Our goal at this step was to achieve a parsimonious data reduction while retaining as much information as our data would permit. Our coding system yielded a binary vector describing the presence or absence of 24 different goal attributes for each goal statement. Recall that goal statements were elicited by six different motivational probes. Thus, for the nine most prevalent codes, we calculated subtotals representing the occurrence of each goal attribute for statements elicited by each motivational theme. For example, this cross-classification allowed us to distinguish among goals about solving money problems, earning more money, or learning to live more frugally. The 54 variables formed by the cells of this cross-classification of nine major goal attribute codes by six motivational themes fully accounted for 74% of all responses. These represented our primary goal attributes. We reduced these 54 variables by conducting a two-stage principal components analysis, first summarizing endorsement rates of codes within each of the six motivational themes (retaining 57% to 89% of total variance in each set), and then combining the 32 first-order components from these six analyses in a single second-order principal components analysis (retaining 60% of variance among the first-order components). Second-order analysis yielded 16 major goal attribute factors. These components are listed in Table 1.

The remaining 15 goal attribute codes were less prevalent, so we simply tallied the total number of times content codes occurred for each individual at a given time of measurement, without subtotaling by eliciting theme. Principal components analysis yielded seven relatively independent components after promax rotation, summarizing 58% of the variance among these 15 codes. These seven subsidiary goal attribute dimensions are also listed in Table 1. Substantively, we think of the primary goal content factors as capturing the individuals' status in broadly shared areas of concern, whereas the subsidiary dimensions reflect more particular concerns that may nonetheless have an important influence on individuals' appraisal of QOL.

2.4. Scoring other domains of QOL appraisal battery

The other three parameters of QOL appraisal were analyzed by a series of principal component analyses to map the items of *sample experiences* to five factors, *standards of comparison* to three factors, and *combinatory algorithms* to seven factors. Generally, principal components with eigenvalues greater than or equal to 1 were retained. Tables 2−4 summarize the total variance accounted for by the retained eigenvalues and the rotated factor loadings of the items. Standardized factor scores were calculated and entered into the analysis. Take the *combinatory algorithms* scale in Table 4 as an example, respondents were prompted "When you answered today, did you think more about…" and they rated the extent to which they thought about "Things that are disappointing to you," "How hard it has been," "Things that make you feel worried," and so on. These three items had high-factor loadings on the first factor that was thus labeled as "Negative Experiences, Feelings, & Worries."

2.5. Changes in QOF appraisal

Because the QOL appraisal subscales were standardized, all subdomains were thus mapped onto a comparable scale of mean zero and unit standard deviation (SD). A respondent with a zero "reacting to recent flare-ups" score, for example, represents an appraisal through recent disease flare-ups at the sample average. Changes in QOL appraisal were thus operationalized as changes in the standardized scores. In principle, the changes in standardized scores can be thought of as changes in effect-size units [17,18], thereby simplifying comparisons made across multiple QOL appraisal domains on arbitrary raw scales. We felt that it would facilitate interpreting the changes in appraisal by considering a set of crude but practical cutoffs. We considered a 0.75 SD change a "moderate" change in appraisal, a 1.0 change a "moderately large" change, and a 1.5+ change a "large" change. The "large" change is conveniently twice as large as the "moderate" change. These cutoffs are more conservative than the conventional effect-size indexes [18] (e.g., 0.80 as a "large"

Table 1
Primary and subsidiary goal content dimensions

	Primary goal content dimensions	Example verbatim response
1	Maintain relationships, accept others, improve outlook	"I want to keep my marriage the same as it is now" (865.6)
2	Solve problems with HIV treatment	"I want to solve my HIV problems" (887.2)
3	Prevent money and housing problems, reduce worries and obligations	"I want to accomplish getting better housing" (578.3)
4	Solve problems related to living situation and work	"I want to solve problems in my living situation, I need to move" (53.4)
5	Address physical and emotional health problems	"I want to prevent or avoid pain, both physical and emotional" (313.4)
6	Learn to live with HIV diagnosis and maintain current treatment arrangements	"I want to accept that I've got to live with HIV and accept as they are" (325.6)
7	Avoid interpersonal and monetary concerns	"I want to avoid getting into fights and arguments" (1849.8)
8	Maintain a positive mood, learn to accept the inevitable	"I want to keep the same my attitude in thinking positively" (2116.9)
9	Accomplish work and financial goals	"I want to be more financially stable" (177.1)
10	Avoid work-related problems	"I don't want anybody at work to know my HIV status" (137.8)
11	Reduce practical and monetary obligations and demands	" I want to reduce responsibilities such as paying credit card bills" (1808.9)
12	Maintain current living situation (vs. address health problems)	"I want to keep my living arrangements with my wife and family the same" (2331.10)
13	Acceptance, resignation to health and mood problems	"I want to accept that nobody lives forever" (783.4)
14	Maintain and accept current work and monetary situation	"I want to accept that I have to work even with my health problems" (2127.7)
15	Acceptance of living conditions, housing, and neighborhood	"I want to keep living in my neighborhood" (2565.6)
16	Concerns about HIV prevention	"I want to prevent/avoid infecting other people with HIV" (209.1)
	Subsidiary goal content dimensions	
1	Outreach and community concerns	"I want to accomplish more in outreach work" (210.1)
2	Social, religious, and discrimination concerns	"I want to turn to God more" (1076.7)
3	Travel and leisure vs. chores	"I want to travel to my country" (1707.7)
4	Independent functioning	"I want to keep the same my independence" (2053.9)
5	Education and self-fulfillment	"I want to accomplish the goal of finishing school" (367.8)
6	Legal and immigration concerns	"I want to solve the problem of my immigration status" (2261.10)
7	Substance use	"My main problem to solve is to break the methadone habit" (781.1)

Selected verbatim responses are identified by two numbers so that (781.1) represent statement number 781 assigned to rater 1.

effect), and we believe that they help track the numerous and complex pattern of splits in the rpart analysis.

2.6. HRQOL discrepancy score

After Rapkin and Schwartz' [1] formulation to examine HRQOL response shift, it was first necessary to derive a discrepancy score in the HRQOL to determine how much the observed score differed from an expected value. Response shift arises when the observed changes in HRQOL scores deviate systematically from the expected HRQOL changes owing to health-related events. A simplistic example illustrates the basic conceptual premises. If a person experiences more symptoms related to HIV/AIDS, and the

Table 2
Factor loadings of quality-of-life appraisal scales in sampling experiences

	1[b]	2	3	4	5
"When you responded today, how much did you…"	14%[a]	14%	13%	11%	9%
Find yourself thinking about the worst moments?	0.78				
Focus on HIV/AIDS?	0.67				
Consider things that you'd only think about for an interview like this?	0.57				
Try to give your first reaction to the questions?		0.77			
Try not to complain too much?		0.74			
Try to communicate the seriousness of your situation?		0.64			
Try to remember everything relevant over the past 3 months?			0.78		
Think about how things have been going over past few days?			0.70		
Emphasize the positive as much as possible?	−0.42		0.51		
Consider your relationships with family/friends?				0.83	0.46
Take into account what your doctor has told you about your health?					
Think about the future?				0.59	0.82
Balance the positives with the negatives?		0.42		0.44	
Recall recent episodes or flare-ups?	0.44		0.43		

Note: loadings <0.40 were omitted.

[a] Variance accounted for by the eigenvalues associated with the latent factors.

[b] Factor 1: "Focusing on worst moments re-illness;" Factor 2: "Formulating responses to manage interview;" Factor 3: "Recalling recent events;" Factor 4: "Considering interactions with family and others;" and Factor 5: "Contemplating the future."

Table 3
Factor analysis and factor loadings on focus of comparisons

"When rating your health and well-being today, how much did you compare yourself to…"	1	2	3
	23%	23%	20%
Others you know who are living now with HIV/AIDS?	0.89		
People whose health does not limit them in any way?	0.89		
Your ideal: your dream of perfect health?		0.82	
The kind of life that you are really working for?		0.82	
A time in your past before you had HIV?			0.88
Most people your age?		0.40	0.63
The way that the people in your life see you?			0.45
The things your doctor told you would happen?			0.41

1. Comparing oneself to others with HIV and with no health limitations; 2. Comparing oneself to personal ideals or desired goals; and 3. Comparing oneself to one's past and to age-related norms.

increased symptoms are expected to reduce HRQOL by 10% (such as the predicted HRQOL change by a statistical model derived from large-scale surveys, controlling for other validated covariates), then an observed increase of 15% suggests response shift. Thus, if change in cognitive appraisal was able to explain these systematic discrepancies, that was indicative of response shift. We decided to focus on the mental component summary (MCS) score of the SF-36v2 rather than the physical component for this demonstration, because our prior preliminary analysis showed that it is more sensitive to response shift [19].

To provide a highly conservative test of response shift, we used an ordinary least square regression to control baseline mental composite score for a wide range of possible predictors, including demographics and personal history (e.g., history of hard drug use and involvement in the criminal justice system), baseline health status, baseline frame of reference, baseline sampling, standards, and combinatory algorithm, change in health status variables, changes in number of self-reported symptoms, and intervening events in care. Details on how these covariates are assessed can be found in Refs. [14,15]. Standardized residual scores controlling these predictors were computed and entered into the rpart analysis [20–22] to determine whether and how changes in cognitive appraisal could be used to explain these discrepancies.

2.7. *Rpart model specifications*

The rpart [20–22] model fitted the standardized residual MCS scores in SF-36v2 [16] with 38 predictors representing the changes in appraisal variables between baseline and 6-months assessment—changes in 16 primary and seven subsidiary goal content dimensions, five predictors on the sampling of experiences, three predictors on standards of comparisons, and seven predictors on the combinatory algorithms. For ease of interpreting the magnitude of response shift, we divided our sample by the MCS discrepancy scores to three categories—40% with the largest positive residuals (deemed "Positive" response shift), 40% with the largest negative residuals ("Negative" response shift), and 20% with residuals close to zero ("No Change").

We followed the general approach in rpart analysis—first grow a complex tree and then prune the tree back by cross-validation [20,21,23–29]. Feldesman [29] is a highly accessible tutorial on the different statistical computations for continuous and categorical outcome variables; it also outlines a few default model specifications in the complex tree: (1) stopping rule for a terminal node (<20 observations), (2) criterion for tree pruning ("cost-complexity parameter," CP = 0.01), (3) validation by 10-fold cross-validation (1-standard error [1-S.E.] rule for pruning by CP), (4) specification of priors (proportional to data counts), and (5)

Table 4
Factor analysis and factor loadings on salience of experiences

	1	2	3	4	5	6	7
"When you answered today, did you think more about…"	14%	12%	10%	9%	8%	8%	7%
Things that are disappointing to you?	0.78						
How hard it has been?	0.65						
Things that make you feel worried?	0.64						
Things that you are proud of?		0.74					
How well you are doing?		0.70					
Things that make you feel calm?		0.70					
Big changes?			0.78				
The way things usually are?			−0.76				
Things settled to your satisfaction?				0.83			
Things that are unfinished?				−0.74			
What is important to others?					0.76		
What is important to you?					−0.77		
Long time concerns?						0.79	
Recent concerns?	0.40					−0.72	
What you do on your own?							0.90
The help you need from other people?	0.46						−0.52

1. Negative experiences, feelings, & worries; 2. Sources of satisfaction; 3. Change vs. routine; 4. Settled vs. unfinished concerns; 5. Things important to others vs. self; 6. Long-time vs. recent concerns; and 7. Independence vs. help from others.

missing data are handled by surrogate splits. Details can be found in textbooks and are omitted here [13,28,30,31].

2.8. Rpart model fit evaluation

Model performance was evaluated by a three-class classification performance metric based on overall error rate in a confusion matrix and also by pairwise area under the Receiver Operating Characteristic curve (AUC under the ROC) analysis [32,33]. Our three-class classification was separated into six binary comparisons *after* the rpart classifier has been carried out. We calculated the AUC on "Positive" vs. "No Change" response shift, "Positive" vs. "Negative" response shift, and so on for all six pairwise ROCs. A single average AUC index was calculated, called the M function [33], to represent the overall model performance. We also entered the same 38 predictors in a multinomial logit model for comparison.

3. Results

3.1. Respondent characteristics

At this time of analysis, 619 individuals were recruited to this study, of which 443 were due for the 6-month assessments and 394 completed them (89%, follow-up data collection ongoing). Table 5 summarizes participant characteristics. Men and women were approximately evenly distributed, with diverse race and ethnicity backgrounds, low socioeconomic status, and an average age of 47.1 and 11.6 years since the identification of HIV.

3.2. Response-shift analysis using rpart

Figure 1a shows the fullest rpart dendrogram, derived by accepting the default settings. The 10-fold cross-validation suggested pruning the tree back to only nine terminal nodes (Fig. 1b). This was based on the 1−S.E. rule [20,21], plotted in Fig. 2, to find the least complex tree within 1 SD of the minimal cross-validation error. The pruned tree showed the lowest cross-validation error, beyond which tree complexity entailed no additional improvement.

Table 5
Participant characteristics

Characteristic	$n = 619$
Sex (% male)	328 (53%)
Age in yr	47.1 (SD = 8.5)
Time since HIV identification (yr)	11.6 (SD = 5.7)
Marital status/domestic partner	409 (66%)
Sexual orientation (% heterosexual)	452 (73%)
Race	
African descent	359 (58%)
Anglo	31 (5%)
Latino	186 (35%)

Abbreviations: SD, standard deviation.

Note: Numbers are persons and percentages unless otherwise noted.

Table 6 shows the confusion matrix of the nine-node tree and the model performance AUC measures of three alternative models. The nine-node tree made 243 correct classifications (62% accuracy, 95% confidence interval: 39−80% by bootstrapping), which was superior to the 36% chance accuracy by the marginal 40−20−40 split. The pairwise AUC indexes show comparable performance between the nine-node rpart tree and the multinomial logit, with an overall AUC of 0.72. The 24-node tree consistently outperforms the pruned tree and the multinomial logit model. However, the cross-validation argued against it because of the low generalizability.

We now discuss the nine terminal nodes in Fig. 1b from left to right by interpreting the distinctions among groups that emerged in this analysis. The first group of 78 individuals in node 1 stood out because they reduced the salience of negative experiences by a moderately large amount. This group tended to have a high prevalence of positive discrepancies (47 out of 78). For the remaining 316 individuals in nodes two through nine, the salience of negative experiences in evaluating HRQOL was either maintained or increased ($\geqslant -1.04$ SD). For succinctness, we interpret nonlarge reduction in splits as roughly maintenance or possible increase. Reduction in salience of negative experiences alone was not sufficient to affect discrepancies in HRQOL. A combination of other cognitive variables comes into play. The second, third, and fourth nodes were distinguished from the rest of the sample based on a moderate reduction in the extent to which they compared themselves to others. Group 4 represented a small subgroup that differed from groups 2 and 3 by a moderately large increase in goals related to solving problems associated with living situations and work. Although three persons displayed negative discrepancies, most of the individuals in this small group 4 demonstrated little or no discrepancy from expected change in psychological well-being. For individuals in nodes 2 and 3, discrepancy was associated with moderate changes in goals related to independent functioning. Group 2 maintained or increased goals related to independence that was associated with predominately positive response shift. Conversely, group 3 markedly reduced goals associated with independence, contributing to more negative psychological well-being than expected. It is noteworthy that groups 2, 3, and 4 were all affected by changes in specific goals related to problems with work or with maintaining independence. Such changes in frame of reference interact with changes in standards of comparison to produce a range of response shifts.

Nodes 5−9 either maintained or increased their tendency to compare themselves with others, as well as the salience of negative feelings and experiences. On the far right node 9, the largest group of 145 individuals identified in this analysis sample, stood out from the others because of maintained or increased concerns about monetary obligations and other external demands. This combination, greater salience of negative feelings and comparison of self to others along with increased demands, was clearly associated with marked

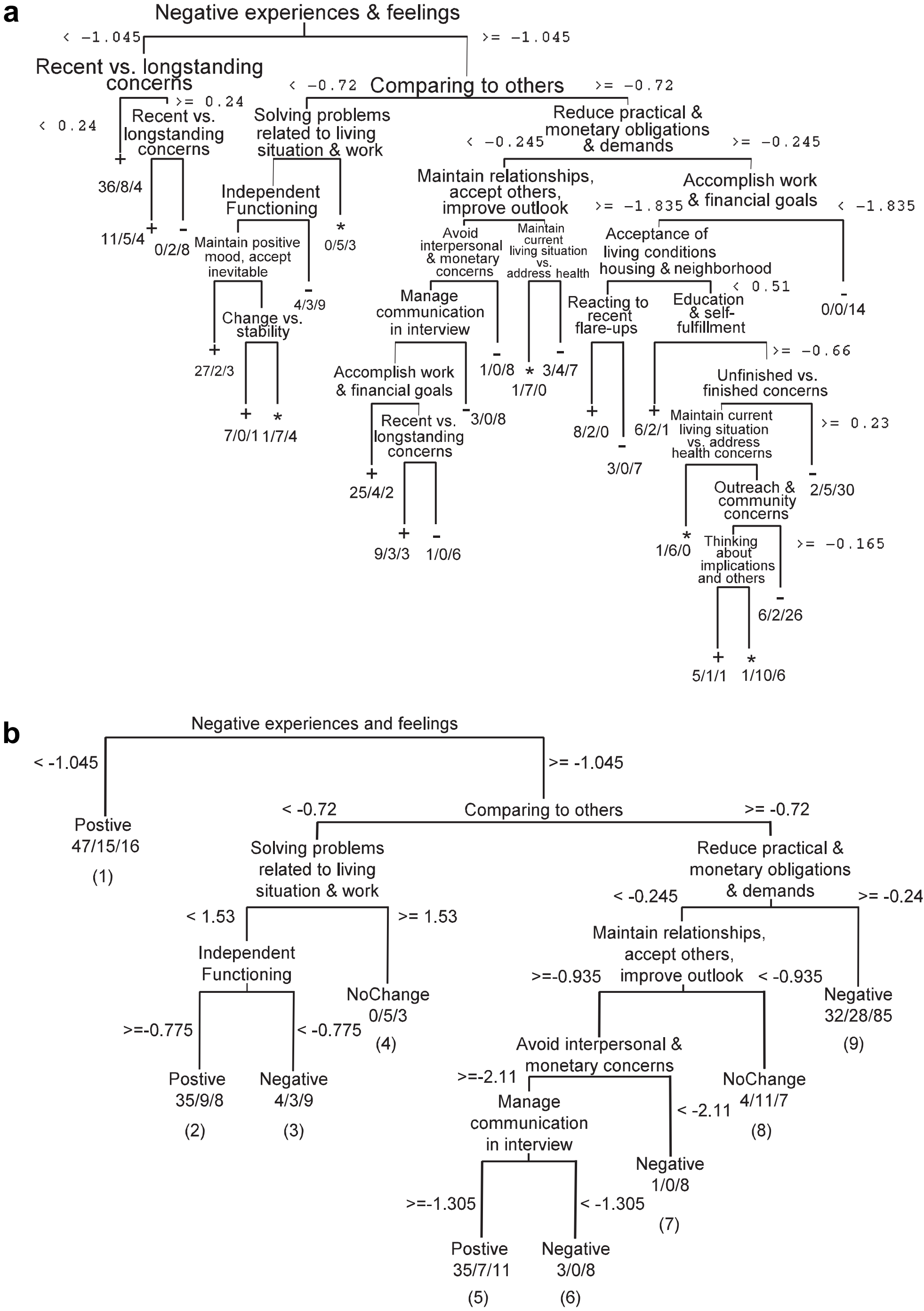

a
Negative experiences & feelings
< -1.045
>= -1.045
Recent vs. longstanding concerns
< 0.24
>= 0.24
+ 36/8/4
Recent vs. longstanding concerns
+ 11/5/4
- 0/2/8
< -0.72
Comparing to others
>= -0.72
Solving problems related to living situation & work
Independent Functioning
* 0/5/3
Maintain positive mood, accept inevitable
- 4/3/9
Change vs. stability
+ 27/2/3
+ 7/0/1
* 1/7/4
Reduce practical & monetary obligations & demands
< -0.245
>= -0.245
Maintain relationships, accept others, improve outlook
>= -1.835
Accomplish work & financial goals
< -1.835
- 0/0/14
Avoid interpersonal & monetary concerns
Maintain current living situation vs. address health
Manage communication in interview
- 1/0/8
* 1/7/0
- 3/4/7
Accomplish work & financial goals
- 3/0/8
Recent vs. longstanding concerns
+ 25/4/2
+ 9/3/3
- 1/0/6
Acceptance of living conditions housing & neighborhood
Reacting to recent flare-ups
< 0.51
Education & self-fulfillment
+ 8/2/0
- 3/0/7
>= -0.66
Unfinished vs. finished concerns
+ 6/2/1
Maintain current living situation vs. address health concerns
>= 0.23
- 2/5/30
Outreach & community concerns
* 1/6/0
Thinking about implications and others
>= -0.165
- 6/2/26
+ 5/1/1
* 1/10/6
b
Negative experiences and feelings
< -1.045
>= -1.045
Postive 47/15/16 (1)
< -0.72
Comparing to others
>= -0.72
Solving problems related to living situation & work
< 1.53
>= 1.53
Independent Functioning
NoChange 0/5/3 (4)
>=-0.775
< -0.775
Postive 35/9/8 (2)
Negative 4/3/9 (3)
Reduce practical & monetary obligations & demands
< -0.245
>= -0.245
Maintain relationships, accept others, improve outlook
>=-0.935
< -0.935
Negative 32/28/85 (9)
Avoid interpersonal & monetary concerns
>=-2.11
< -2.11
NoChange 4/11/7 (8)
Manage communication in interview
Negative 1/0/8 (7)
>=-1.305
< -1.305
Postive 35/7/11 (5)
Negative 3/0/8 (6)

Fig. 1. Full recursive partitioning tree with 24 terminal nodes (a) and the pruned tree with nine terminal nodes (b). To minimized clutter in (a), the most prevalent outcomes in the terminal nodes are represented in symbols, for positive discrepancy (marked with the + sign), negative discrepancy (−), and no discrepancy (*). Many splitting criteria in (a) are omitted.

negative discrepancies in reported QOF. However, for individuals without concerns about reducing demands and obligations, other factors came into play. Node 8 represented a group that had moderately high reduction in goals on maintaining relationships by accepting others and improving their own outlook. This group tended to report changes in psychological well-being close to values predicted by baseline factors, health changes, and events in care.

Individuals in the remaining groups 5, 6, and 7 all tended to increase or sustain their concerns about maintaining relationships and achieving a positive outlook. Again moving in from the right, node 7 contained a preponderance of individuals with negative response shift. Interestingly, this group reported a marked decrease in goals related to preventing or avoiding interpersonal and monetary concerns. Conceivably, these individuals wanted to stave off certain

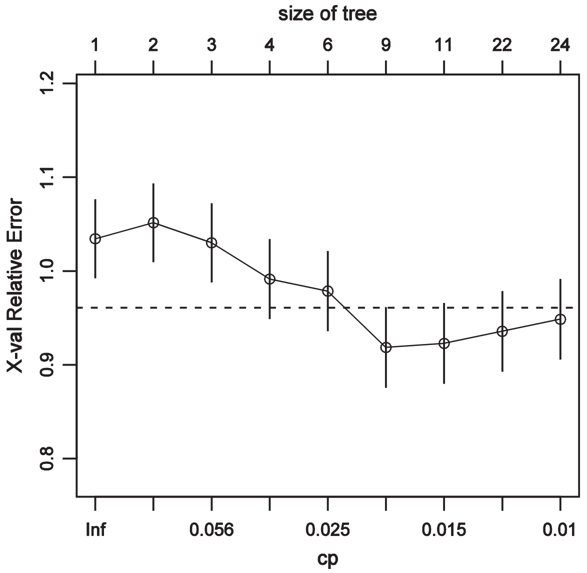


Fig. 2. Graphical output of the plotcp() command in recursive partitioning. The horizontal dotted line represents the 1-standard error rule—the cutoff cross-validation error statistic at 1 standard deviation above the minimal cross-validation error. The tree with nine terminal nodes is considered the desired size for pruning because it entails the lowest cross-validation before additional complexity in the tree is accompanied by higher cross-validation errors.

problems at baseline but later realized that this was untenable by 6 months, thus contributing to a negative response shift.

Nodes 5 and 6 share many features, including the salience of negative experiences, a tendency to compare oneself to others, goals to maintain relationships, improve one's outlook, and avoid interpersonal and monetary problems, but not to reduce obligations or demands. Most of these individuals, in node 5, tended to report more positive psychological well-being than expected as a result of efforts to manage communications during the interview. This factor, from the sampling experience domain reflects individuals' efforts to edit their responses by not complaining too much, by giving their first reaction, and by trying to convey the seriousness of their situation. Increasing or sustaining this response set was associated with positive discrepancies in well-being. Alternatively, individuals in node 6 had reduced or abandoned efforts to manage communication during the interview. Most of the individuals in node 6 demonstrated negative response shift.

4. Discussion

The rpart-derived model was useful in identifying aspects of response shift that was hard to detect through linear analysis. Rpart performed equally well as a multinomial logistic regression of the same predictors. Rpart provided a straightforward method that yielded more clinically interpretable results for identifying subgroups of response shift that appeared to be mostly influenced by changes in emotion (e.g., "negative experiences and feelings") and subjective norms ("comparing with others"). Invoking comparisons with others played an important role in explaining discrepancies in SF36v2 change that only became apparent when examined in conjunction with the salience of negative events. Thereafter, frames of reference and individual concerns came into play in response shift, including five of the 16 primary goal factors. Our findings also shed light on the potential for individuals to sample and report experiences that affect their HRQOL selectively, to manage communication during the interview. Permitting the interplay of cognitive change variables provides additional, complementary information about QOL response shift.

The present findings have bearings on application of QOL measures and on understanding and meeting patients' needs. Individuals continually encounter new challenges and new opportunities, and factor these into how they self-evaluate their well-being. Individuals living with a chronic, life-threatening illness encounter many such challenges and the stakes are high. An interpersonal conflict may interfere with an important source of social support; monetary or housing problems may strain an individual's limited resources; and challenges to independence may invoke personal fears of premature debilitation and mortality. These challenges may be even greater in the Medicaid HIV/AIDS population. Additionally, we show that

Table 6
Classification and misclassification results of the pruned rpart tree with nine terminal nodes and model performance indexes for three alternative models

		Predicted response shift					
		Positive	No change	Negative	Total		
Residual changes in HRQOL	Positive	117	4	40	161		
	No change	31	16	31	78		
	Negative	35	10	110	110		
	Total	183	30	181	394		
	Pairwise AUC						
	1	2	3	4	5	6	Overall AUC
Rpart 24-node tree	0.83	0.90	0.87	0.83	0.90	0.87	0.86
Rpart 9-node tree	0.70	0.77	0.69	0.70	0.76	0.69	0.72
Multinomial logit	0.71	0.75	0.71	0.71	0.75	0.71	0.72

Abbreviations: Rpart, recursive partitioning; HRQOL, health-related quality of life; AUC, area under the ROC curve.

1. No change vs. positive; 2. Negative vs. positive; 3. Negative vs. no change; 4. Positive vs. no change; 5. Positive vs. negative; and 6. No change vs. negative.

some individuals adapt to challenges by altering their cognitions, by relinquishing goals or modifying expectations for what they are seeking (e.g., node 9 by ''reducing practical & monetary obligations/demands'' and node 7 by ''avoiding interpersonal & monetary concerns''). The distinction between nodes 7 and 9 is subtle, with node 9 emphasizing more on practical concerns, such as reducing the financial obligations of paying bills. Individuals may selectively disengage from situations that they can no longer manage. These processes necessarily play out overtime. It is not surprising that individuals' cognitive criteria for the appraisal of psychological well-being and distress are quite fluid.

Cognitive assessment provides a way to take these wide variations in QOL appraisal into account [1]. We can use these methods to control response-shift effects in evaluations of programs or treatments. For example, we might observe improvement in an individual's emotional well-being if we take into account that they are presently engaged in solving housing problems and in boosting independence (e.g., node 2). Similarly, apparent reduction in emotional well-being might be reinterpreted in light of an individual's selective emphasis on reducing practical and monetary demands (e.g., node 9). More fundamentally, it may be important to interpret the impact of disease and treatment on measures of cognitive appraisal. As our analysis demonstrates, there is a complex interplay among measures of appraisal and QOF. It is important to understand when and how increased contact with the health system is associated with a sense of greater dependence, and when it is associated with increasing expectations and standards for self-evaluation. Change in the content and process of cognitive appraisal is a worthy patient-reported outcome domain in its own right.

Our results support the Rapkin and Schwartz model [1], in that cognitive variables helped to account for substantial HRQOL response shift in ways that were interpretable and consistent. However, several methodologic challenges remain. The QOL appraisal battery generates considerable, detailed descriptive data about the appraisal process. It is quite challenging to operationalize the process of describing the intermediaries of response shift, as we have attempted. There are inherent problems in CART methods [34] that originate from the fact that one predictor may win a particular split by only a small margin. This makes such splits somewhat arbitrary; errors cascade into subsequent splits, highlighting the importance of tree pruning by cross-validation [34]. New methodologic developments are available [35–38] and may be helpful in future research on response shift. We hope that our study will prompt further theoretical and empirical work to improve on the description of the response-shift phenomena.

Acknowledgments

The authors acknowledge funding from the New York State Department of Health AIDS Institute (US Health Resources and Services Administration grant: 2X07 HA 0025-17) (B.R.) and the Weill Cornell Medical College Clinical and Translational Science Award (NIH UL1-RR024996) (Y.L.).

References

[1] Rapkin B, Schwartz CE. Toward a theoretical model of quality-of-life appraisal: implications of findings from studies of response shift. Health Qual Life Outcomes 2004;2:14.

[2] Rapkin B. Personal goals and response shifts: understanding the impact of illness and events on the quality of life of people living with AIDS. In: Schwartz CA, Sprangers MAG, editors. Adaptation to changing health: response shift in quality-of-life research. Washington, DC: American Psychological Association; 2000. p. 53–71.

[3] Schwartz CE, Sprangers MAG. Methodological approaches for assessing response shift in longitudinal quality of life research. Soc Sci Med 1999;48:1531–48.

[4] Sprangers MAG, Schwartz CE. Integrating response shift into health-related quality-of-life research: a theoretical model. Soc Sci Med 1999;48:1507–15.

[5] Ubel PA, Loewenstein G, Jepson C. Whose quality of life? A commentary on exploring discrepancies between health state evaluations of patients and the general public. Qual Life Res 2003;12: 599–607.

[6] Ubel PA, Loewenstein G, Schwarz N, Smith D. Misimagining the unimaginable: the disability paradox and health care decision making. Health Psychol 2005;24:S57–62.

[7] Sackett DL, Torrance GW. The utility of different health states as perceived by the general public. J Chronic Dis 1978;31:697–704.

[8] Schwartz CE, Andresen EM, Nosek MA, Krahn GL. Response shift theory: important implications for measuring quality of life in people with disability. Arch Phys Med Rehabil 2007;88:529–36.

[9] Weiden P, Rapkin B, Mott T, Zygmunt A, Goldman D, Horvitz-Lennon M, et al. Rating of medication influences (ROMI) scale in schizophrenia. Schizophr Bull 1994;20:297–310.

[10] O'Boyle CA, McGee H, Hickey A, O'Malley K, Joyce CR. Individual quality of life in patients undergoing hip replacement. Lancet 1992;339:1088–91.

[11] McGee H, O'Boyle CA, Hickey A, O'Malley K, Joyce CRB. Assessing the quality of life of the individual: the SEIQoL with a healthy and a gastroenterology unit population. Psychol Med 1991;21:749–59.

[12] Sprangers MAG, van Dam F, Broersen J, Lodder L, Wever L, Visser MRM, et al. Revealing response shift in longitudinal research on fatigue: The use of the then-test approach. Acta Oncol 1999;38: 709–18.

[13] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. New York: Chapman Hall/CRC; 1984.

[14] Rapkin B, Weiss E, Chhabra R, Ryniker L, Patel S, Carness J, et al. Beyond satisfaction: using the dynamics of care assessment to better understand patients' experiences in care. Health Qual Life Outcomes 2008;6:20.

[15] Patel S, Weiss E, Chhabra R, Ryniker L, Adsuar R, Carness J, et al. The Events in Care Screening Questionnaire (ECSQ): a new tool to identify needs and concerns of people with HIV/AIDS. AIDS Patient Care STDS 2008;22:381–93.

[16] Ware JE, Kosinski MA, Dewey JE. How to score version 2 of the SF-36® health survey (standard & acute forms). Lincoln, RI: Quality Metric Incorporated; 2000.

[17] Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.

[18] Cohen J. A power primer. Psychol Bull 1992;112:155–9.

[19] Li Y, Rapkin B. HIV/AIDS patients' quality of life appraisal depends on their personal meaning of quality of life and frame of reference. Qual Life Res 2006;15:A-36. Available at. http://www.springerlink.com/content/h48815020815g7w6/fulltext.pdf. Accessed January 15, 2009.

[20] Therneau TM, Atkinson EJ. An introduction to recursive partitioning using the Rpart routines. Rochester, MN: Mayo Foundation; 1997.

[21] Venables WN, Ripley BD. Modern applied statistics with S. New York: Springer Science+Business Media; 2002.

[22] R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: The R Foundation for Statistical Computing; 2008.

[23] Martin MA, Meyricke R, O'Neill T, Roberts S. Mastectomy or breast conserving surgery? Factors affecting type of surgical treatment for breast cancer—a classification tree approach. BMC Cancer 2006;6:98.

[24] Gruenewald TL, Mroczek DK, Ryff CD, Singer BH. Diverse pathways to positive and negative affect in adulthood and later life: an integrative approach using recursive partitioning. Dev Psychol 2008;44:330−43.

[25] Radespiel-Troger M, Rabenstein T, Schneider HT, Lausen B. Comparison of tree-based methods for prognostic stratification of survival data. Artif Intell Med 2003;28:323−41.

[26] Sedrakyan A, Zhang H, Treasure T, Krumholz HM. Recursive partitioning-based preoperative risk stratification for atrial fibrillation after coronary artery bypass surgery. Am Heart J 2006;151:720−4.

[27] Nicolucci A, Carinci F, Ciampi A. Stratifying patients at risk of diabetic complications: an integrated look at clinical, socioeconomic, and care-related factors. SID-AMD Italian Study Group for the Implementation of the St. Vincent Declaration. Diabetes Care 1998;21:1439−44.

[28] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2001.

[29] Feldesman MR. Classification trees as an alternative to linear discriminant analysis. Am J Phys Anthropol 2002;119:257−75.

[30] Zhang H, Singer B. Recursive partitioning in the health sciences. New York: Springer-Verlag; 1999.

[31] Clark LA, Pregibon D. Tree-based models. In: Chambers JM, Hastie TJ, editors. Statistical models in S. New York: Chapman and Hall/CRC; 1991.

[32] Patel AC, Markey MK. Comparison of three-class classification performance metrices: a case study in breast cancer CAD. In: Eckstein MP, Jiang Y, editors. Proceedings of the SPIE, medical imaging 2005: image perception, observer performance, and technology assessment. Bellingham, WA: SPIE; 2005. p. 581−9.

[33] Hand DJ, Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. Mach Learn 2001;45:171−86.

[34] Radespiel-Troger M, Hothorn T, Pfahlberg AB, Gefeller O. Re: Applying recursive partitioning to a prospective study of factors associated with adherence to mammography screening guidelines. Am J Epidemiol 2006;164:400−1.

[35] Breiman L, Cutler A. Random forest. Mach Learn 2001;45:5−32.

[36] Breiman L. Bagging predictors. Mach Learn 1996;24:123−40.

[37] Hothorn T, Lausen B, Benner A, Radespiel-Troger M. Bagging survival trees. Stat Med 2004;23:77−91.

[38] Hothorn T, Hornik K, Zeileis A. Party: a laboratory for recursive part(y)itioning. Available at. www.r-project.org 2006; Accessed January 20, 2009.