Family Medicine and Community Health

Developing and testing an automated qualitative assistant (AQUA) to support qualitative analysis

Robert P Lennon ,[1] Robbie Fraleigh,[2] Lauren J Van Scoy,[3] Aparna Keshaviah,[4] Xindi C Hu,[4] Bethany L Snyder,[5] Erin L Miller,[1] William A Calo,[6] Aleksandra E Zgierska,[1] Christopher Griffin[2]

To cite: Lennon RP, Fraleigh R, Van Scoy LJ, *et al*. Developing and testing an automated qualitative assistant (AQUA) to support qualitative analysis. *Fam Med Com Health* 2021;**9**:e001287. doi:10.1136/fmch-2021-001287

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/fmch-2021-001287).

[1]Family and Community Medicine, Penn State Health Milton S. Hershey Medical Center, Hershey, Pennsylvania, USA
[2]Applied Research Laboratory, Pennsylvania State University, University Park, Pennsylvania, USA
[3]Internal Medicine, Penn State Health Milton S. Hershey Medical Center, Hershey, Pennsylvania, USA
[4]Mathematica Policy Research Inc, Princeton, New Jersey, USA
[5]Center for Community Health Integration, Case Western Reserve University, Cleveland, Ohio, USA
[6]Public Health Services, Penn State Health Milton S. Hershey Medical Center, Hershey, Pennsylvania, USA

Correspondence to
Dr Robert P Lennon; rlennon@pennstatehealth.psu.edu

ABSTRACT

Qualitative research remains underused, in part due to the time and cost of annotating qualitative data (coding). Artificial intelligence (AI) has been suggested as a means to reduce those burdens, and has been used in exploratory studies to reduce the burden of coding. However, methods to date use AI analytical techniques that lack transparency, potentially limiting acceptance of results. We developed an automated qualitative assistant (AQUA) using a semiclassical approach, replacing Latent Semantic Indexing/Latent Dirichlet Allocation with a more transparent graph-theoretic topic extraction and clustering method. Applied to a large dataset of free-text survey responses, AQUA generated unsupervised topic categories and circle hierarchical representations of free-text responses, enabling rapid interpretation of data. When tasked with coding a subset of free-text data into user-defined qualitative categories, AQUA demonstrated intercoder reliability in several multicategory combinations with a Cohen's kappa comparable to human coders (0.62–0.72), enabling researchers to automate coding on those categories for the entire dataset. The aim of this manuscript is to describe pertinent components of best practices of AI/machine learning (ML)-assisted qualitative methods, illustrating how primary care researchers may use AQUA to rapidly and accurately code large text datasets. The contribution of this article is providing guidance that should increase AI/ML transparency and reproducibility.

INTRODUCTION

Despite its value in studying complex public health issues, qualitative research remains underused.[1] In part this stems from the time and cost of annotating data in a qualitative study, a process known as coding.[2] There may also be a desire to publish quantitative data when it is available instead of waiting for the qualitative component to be completed,[3] and, for time-sensitive research questions like those related to COVID-19 behaviours, the time delay to complete traditional qualitative research may circumvent meaningful contributions to public health crises.

Computer-Assisted Qualitative Data Analysis Software (CAQDAS) is commonly used to assist researchers with the management, organisation, and analysis of qualitative data.[4] These software programs began as mechanisms to better organise and code data, and now include analytic tools such as word frequencies, word clustering, sentiment analysis and thematic analysis. All of these features help researchers construct themes from large datasets, but still require manual coding of data within the software package. The process of coding itself remains a time consuming, labour intense process.

Natural language processing (NLP) has been used to code qualitative data in exploratory mixed methods research.[5 6] Guetterman *et al* found NLP coding to be time-efficient and comparable to human coders in identifying major themes, but lacking in the ability to identify nuances.[5] NLP using latent Dirichlet allocation (LDA) as an initial modelling technique was able to generate topic categories from which the researcher identified overall theme sets similar to traditional methods.[7] Modern unsupervised NLP (especially for topic extraction or document classification) has extended beyond the linear algebraic techniques like Latent Semantic Indexing (LSI)[8] or probabilistic techniques like LDA.[9] More recently, Chang *et al* used human thematic analysis to inform NLP algorithms to evaluate clinical records to identify meta-inferences about barriers related to rapid adoption of virtual medicine visits during COVID-19, and separately to evaluate short text survey responses.[6]

These approaches have been extended for time-varying topic analysis of text corpi, including the use of Tensor decomposition methods by Lowe and Berry.[10] Current 'best in class' approaches use Transformer approaches,[11] that is, a deep neural network

approach to language analysis. These techniques can be extended for use in time-varying topic analysis, for example, for Twitter analysis.[12] Techniques using Topic Modelling and Word2Vec produced similar outcomes to traditional qualitative methods in a proof-of-concept application in public health research.[13] However, the challenge with transformer methods is that, even more than LDA, the approach is fundamentally opaque (as is the case with most deep learning techniques) and consequently may be subject to slow uptake by the qualitative science community. Indeed, the lack of transparency and poor reproducibility of artificial intelligence (AI)/machine learning (ML) results have led to calls for best practice guidance in AI/ML research.[14]

We developed an AI/ML platform to augment qualitative analysis by automating components of qualitative coding—the time-intensive process of matching specific qualitative input into response categories developed by the research team—that avoids the opacity of LDA and Transformer approaches. We further integrated visual analytics into the platform to generate useful, visually appealing data displays to facilitate rapid data exploration and knowledge discovery when analysing large datasets by reducing dimensionality[15] and showing hierarchies within data.[16–18] The objective of this paper is to describe a model methodology for primary care researchers to use our automated qualitative assistant (AQUA) to augment qualitative coding of large datasets in an effort to broaden the feasibility of large-scale qualitative research.

Qualitative methods for AQUA application

A detailed review of qualitative analytic methods is beyond the scope of this paper. AQUA may be integrated into qualitative design at two stages of analysis. In early analysis, AQUA enables researchers to conduct rapid thematic analysis of large free-text datasets and generate visually interpretable outputs. After human coders analyse a subset of a large qualitative dataset, AQUA may be used to code some thematic categories across the remaining dataset, markedly increasing the scope of analysis a given team may complete. AQUA is designed to analyse free-text answers to survey questions. Careful question design and data collection methods will improve AQUA's accuracy. An a priori interpretive framework is necessary to maintain the integrity of the qualitative analysis, and care is needed when reporting AQUA-generated results to avoid over-reach and improve generalisability.

Question design and data collection

AQUA capitalises on the epistemological compatibility between text mining and qualitative research.[19] Human-generated text is rife with idiom, non-standard expressions and jargon. AI/ML that works beautifully in a sterile environment may not work when confronted with the gritty reality of human experience.[20] Researchers must, therefore, carefully construct qualitative questions to minimise idiosyncrasies without compromising the goal of open-ended responses. We recommend that draft survey questions be refined using at least 2 rounds of cognitive interviewing procedures using the think-aloud technique,[21 22] followed by pilot testing on a sample of participants from the desired study populations. Throughout this iterative improvement process, questions should be refined to improve answers' qualitative sensibility and linguistic harmony. Qualitative sensibility ensures that the responses are indeed answering your questions. Linguistic harmony improves AQUA's ability to properly categorise responses. For population samples markedly different from the researchers, we recommend employing population sample focus groups, which have been used with success in cross-cultural and cross-linguistic analysis.[23]

Results interpretation and reporting

We compared coding between a human coding team and AI/ML algorithms by comparing the AI/ML-human intercoder reliability (ICR) to the intrateam ICR of human coders. AI/ML-human ICRs which are substantially lower than the human intra-team ICR indicate that the given data is not easily matched to given categories, and is thus not amenable to automated analysis. ICR is commonly measured using Cohen's kappa or Krippendorff's alpha.[24] For simplicity, where it applies we recommend Cohen's kappa. While there is not universal agreement on a minimum acceptable ICR to indicate clinical utility, it is reasonable to use the interpretation rubric developed by Landis and Koch[25]: values <0 = disagreement, between 0 and 0.20=slight, 0.21–0.40=fair, 0.41–0.60=moderate, 0.61–0.80=substantial and 0.81–1=nearly perfect agreement. Researchers should select a minimum ICR based on the intended use of anticipated results. For example, if using AQUA to code data in a grounded theory study to develop a theory with immediate clinical implications (ie, vaccine distribution), researchers might require ICRs indicating near perfect agreement and set an acceptable ICR cut-off of ≥0.81. Researchers looking to better understand a given populations' lived experience using a phenomenology design might not wish to miss potential areas of exploration, and therefore include ICRs that indicate substantial, or even moderate agreement, with acceptable ICR cutoffs set at ≥0.61 or ≥0.41, respectively.

Because AI/ML techniques are able to analyse very large sets of data, including dozens of analytic categories, the AI/ML output can include not only single category comparisons, but dozens of topic clusters, all with a wide range of ICRs. To maximise generalisability and reproducibility, it is incumbent on researchers to clearly identify a priori ICR cut-offs and category selection requirements, and when interpreting results avoid any temptation to select category topics simply to increase ICR, or include desired topics by lowering ICR targets post hoc.[26]

Researchers must clearly report their a priori interpretation frameworks. If unanticipated results are found that suggest a new direction for study that fall outside this framework, it is reasonable to report these results, with the caveat that they must be identified as a post hoc result, which may be less generalisable. For example, suppose

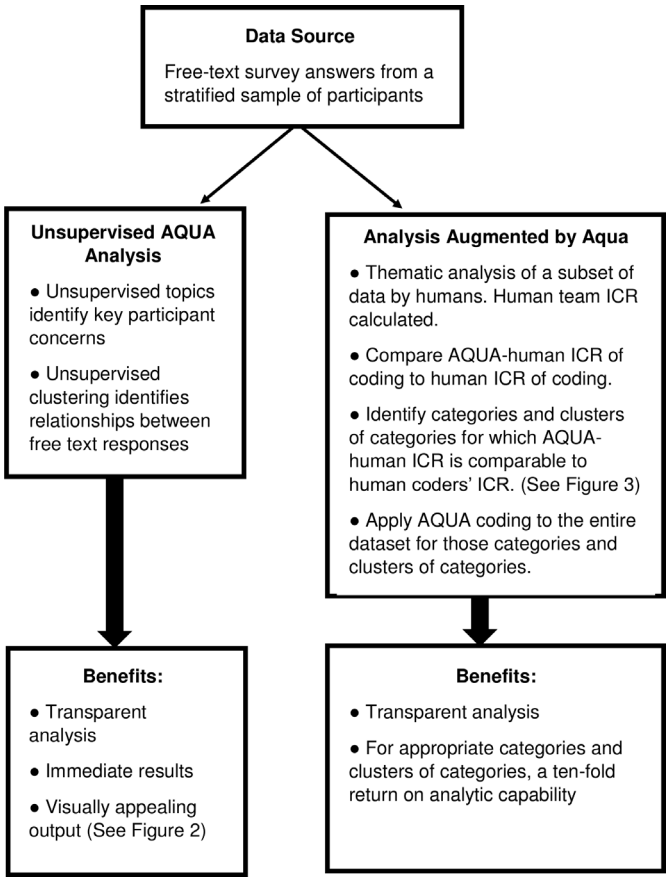

Figure 1 Procedural diagram for the application of AQUA to free-text data used in the Illustration. AQUA, automated qualitative assistant.

the researchers using the grounded theory design above, with an acceptable ICR cut-off of 0.70, found an interesting coding outcome with an ICR of 0.60. Because that falls outside their a priori framework, they must reject that outcome in their primary results. It would be appropriate, however, to comment that while rejected for this work, the moderate agreement found indicates an area that warrants further study.

Illustrating the application of AQUA

To illustrate the utility of AQUA to primary care researchers analysing large text datasets, we present an exemplar study in which AQUA was used to code free-text responses to a survey about public health recommendations related to COVID-19.[27 28] Figure 1 provides an overview of how AQUA was integrated into qualitative analysis to provide immediate, usable outputs and then enable researchers to code elements from the entire dataset.

Data source

The data source was 3148 free-text responses from 538 participants (stratified from 5948 total respondents) who completed a survey to explore the role of trust within information sources related to COVID-19.[27] Six human coders analysed the data using traditional inductive thematic analysis,[29] generating a codebook which identified 11 qualitative categories and 72 subcategories (categories).

Data analysis

Early unsupervised analysis

AQUA uses two methods to code the raw data using this codebook: a semiclassical approach that replaces LSI/LDA with a graphtheoretic topic extraction and clustering method[12 30] and a more modern transformer method based on BERT-based solutions and top2vec.[31] The graph theoretic method is developed in the spirit of Miller's parsimonious topic models[32] but with the Bayesian Information Criterion for determining optimal topic clustering replaced by a maximum modularity (ie, spectral[33]) clustering.[34] We choose these two methods because (i) BERT and top2vec based methods have already been shown to outperform LDA in information theoretic terms while (2) graph theoretic methods lend themselves to visualisation, which is an important element of interpreting data.

The unsupervised clustering approach used is a variation on both LSI[8] and Spectral Clustering,[34] using a maximum modularity subroutine that eliminates the requirement that users choose the number of free-form text response clusters a priori. Responses are clustered into groups with similar linguistic features by creating a response similarity graph and then using maximum modularity clustering to find 'response communities' within the graph. Bags of words for each automatically generated response cluster are computed using a word assignment model that minimises mutual information between the bags of words (subject to some constraints). The dimension of vocabulary space is first reduced using a non-linear dimensional reduction method. Specifically, a trimmed term-response matrix is formed by removing common and non-key words. A term-graph is then formed and maximum modularity clustering is used to find an orthogonal topic basis. Graph edges are words comentioned in a response. This process is similar to principal components analysis[35] (or LSI[8]) but the projection is mediated by the graph clustering step, which handles non-linearity in a similar manner to manifold learning.[36] Hierarchical clustering is then performed by iteratively executing the unsupervised clustering procedure for each individual response community so that the linguistic diversity of each subtopic can be preserved.

Analysis of AQUA's coding accuracy

Supervised word/phrase assignment of words to response clusters (eg, human or machine-created code categories) is accomplished by solving a linear assignment problem[37] of words/phrases to preclustered response groups. The assignment problem minimises a linearised version of the mutual information between text clusters (in word/phrase space) resulting in a parsimonious weighted assignment of words/phrases to responses. These words/phrases characterise the underlying language within the response groups (code categories). The weighted bag of

words were then used to assign new text to one of the pre-existing categories.

Unlike traditional machine-learning processes where the algorithm trains on 90% of the dataset and tests on 10% of the dataset, we trained on much smaller dataset subsets. This is because human coding is time-expensive and we sought to develop algorithmic robustness to perform well using small training datasets. AQUA was tasked with coding the raw data. Item codes were assigned using cosine-similarity on text in the response and the weighted bag of words identified during the supervised learning process. In using this approach, we relied on the fact that text is highly separable[38] and adapted the graph-theoretic methods that underlie our initial methods as a graph-based manifold regularisation approach[39] (in a semisupervised context).

Summary of analysis

Unsupervised clustering and topic selection were used to identify areas of importance to survey respondents and relationships between responses. Highly accurate coding categories by AQUA were identified for further automated analysis. To evaluate the accuracy of AQUA's coding, a human coding team first coded the same data using the same codebook. The human team had an intrateam ICR using Cohen's kappa of ≥0.65 among six human coders (two coders per response). For an AQUA-coded topic to be accepted, we set the AQUA-human ICR cut-off at ≥0.65 (at least as good as the intrahuman team ICR). Categories and clusters of categories with AQUA-human ICRs ≥0.65 were deemed suitable for AQUA coding of the entire dataset (over 35 000 free-text responses from 5948 respondents[28]).

RESULTS

AQUA generated a circle hierarchy of free-text responses and also seven unsupervised topics (figure 2). This enabled rapid assessments of key issues important to survey participants, and also identified relationships between responses in an easy-to-read circle hierarchy.

AQUA's kappa for coding all categories was low (kappa ~0.45), reflecting the challenge of automated analysis of diverse language. However, for several three-category combinations (with less linguistic diversity), AQUA performed comparably to human coders, with an ICR kappa range of 0.66 to 0.72 based on test-train split (table 1; see online supplemental material for all three-category results.) AQUA may appropriately be applied to the entire dataset for categories and groups of categories with AQUA-human ICRs ≥0.65. Figure 3 shows the relationship between automatically generated message clusters (topics) and manually coded categories.

A secondary result was the time spent by AQUA (including human interpretation of AQUA's results) compared with the time spent by human coders. The human coding team spent approximately 30 person-hours to complete their traditional coding. All results

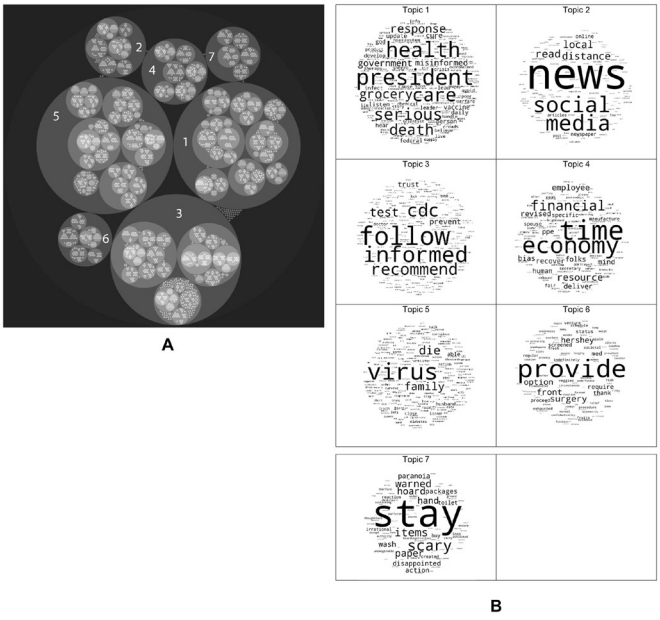

Figure 2 Unsupervised clustering (circle hierarchy) of survey responses. (A) Circles represent a linguistic community or topic. Nested circles represent hierarchical organisation. The smallest circles are individual survey responses. The seven parent unsupervised topics are labeled. (B) Parent unsupervised topics.

generated by AQUA (including human interpretation) took approximately 5 hours.

DISCUSSION

This study demonstrates the feasibility of using AI methods to augment certain kinds of qualitative research. Like current exploratory NLP applications, AQUA offers coding capability beyond was is available in current CAQDAS applications. The primary benefit of the AQUA platform over current exploratory NLP applications[5 6] is the transparency of the graph theoretic approach, which avoids the opacity inherent to NLP based on LDA or Transformer methods.[11] Results of our study suggest that graph-theoretic methods are well suited to augment qualitative researchers during coding, and when integrated into a data visualisation and statistical comparison programme offers qualitative researchers a powerful assistant to enable rapid, rigorous, analysis of large, qualitative datasets.

Unsupervised applications of AQUA offer immediate topic generation and a circle hierarchy of responses which enables rapid analysis without time-intensive human coding. If human coding is completed on a subset of data, for coding categories or clusters of categories in which AQUA's ICR meets a priori ICR cutoffs, AQUA may then be applied to those categories across the entire dataset. In the illustration, human coding time to code 3148 free-text responses from a stratified samples of 538 respondents enabled AQUA to code over 35 000 free-text responses from the entire pool of 5948 respondents

Table 1 Kappa values evaluating the inter-rater reliability (agreement) between human coders and supervised training of three-topic training models

	Train percentage			
Topics	**0.2**	**0.4**	**0.6**	**0.8**
124	0.67 (±0.01)	0.70 (±0.02)	0.68 (±0.03)	0.70 (±0.04)
135	0.66 (±0.02)	0.66 (±0.02)	0.67 (±0.05)	0.67 (±0.06)
145	0.67 (±0.02)	0.71 (±0.02)	0.70 (±0.02)	0.72 (±0.02)
245	0.67 (±0.04)	0.69 (±0.03)	0.71 (±0.03)	0.72 (±0.06)
257	0.68 (±0.02)	0.70 (±0.01)	0.71 (±0.02)	0.70 (±0.02)

Topic labels are: (1) distrust, (2) media messaging, (3) trusted sources of information, (4) personal medical concerns, (5) family concerns, (6) societal concerns, (7) barriers to recommendations, (8) no worries, (9) other Broad.

for certain categories and category clusters—a 10-fold increase in coding power.

Train-test splits of 20%–80% appear adequate to achieve ICR >0.7 for certain combinations of topics. Variance of kappas by subtopic is not unexpected; categories of inquiry that themselves use complex linguistic patterns, or whose associated raw text to be coded to that category use complex linguistic patterns, will be more challenging for AI/ML algorithms to match human ability. Our kappa range of 0.66–0.72 represents 'substantial agreement' under Landis and Koch's rubric.[25] For particularly challenging datasets or discovery applications, lower ICRs may be acceptable. For example, a combination of topics with an ICR of 0.45 (moderate agreement) might not suggest changes to clinical practice, but might serve as a valuable indicator of where to direct future research or conduct additional human analysis. While the generalisability of results may vary with the 'acceptable' ICR, as noted above, the integrity of qualitative analysis may be maintained by reliance on systematic application of an a priori interpretive framework.[26]

We also caution against researchers modifying their category selection to boost kappas using AQUA (or any other AI/ML technique). While it is certainly appropriate to be precise in category descriptions, it is incumbent on the researcher to match the technique to the data, and not vice versa. Some categories or their answers may simply not be amenable to augmentation using this technique. Further, it is important to appreciate that coding accuracy is necessary but insufficient for interpretation. While qualitative models will continue to evolve, incorporating increasingly sophisticated AI/ML algorithms, the heart of interpretation and sensibility lies in human interpretation.

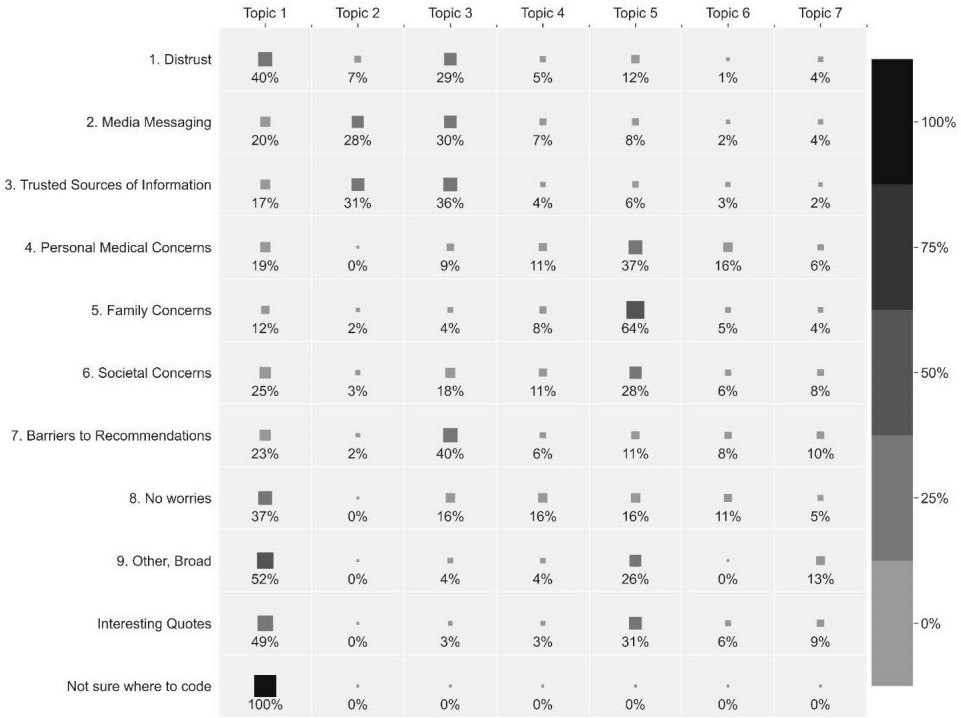

Figure 3 Human-coded topics (rows) are compared with unsupervised topic communities (columns). Each row depicts the distribution of responses for each humancoded topic across unsupervised topic communities. The heat chart to the right shows the gradient color scheme for percent-agreement (darker is better agreement).

Contribution to family medicine and community health research and future directions

AQUA's use of a graph-theoretic approach advances the theory of AI/ML in clinical research. Our methods advance researchers' approach to AI/ML applications in qualitative research, and are particularly useful for very large datasets on which some level of qualitative analysis is already completed. Once a set of category coding is validated on a sample, AQUA enables researchers to rapidly conduct mixed methods analysis on the entirety of the dataset.

In addition to facilitating greater analysis of other existing survey data, the platform and methods may be applied to any set of text data, such as social media posts. This offers researchers a method of rapid and in-depth assessment of any subject of interest and moment, which may improve scientific recommendations, in turn improving policy decisions. The next steps in advancing the AQUA platform are to verify its accuracy across other existing free-text surveys and social media platforms.

Limitations

Limitations of AQUA include that its augmentation focuses on supporting free-text analysis, which may not translate to other areas of qualitative research. Our testing of the AQUA approach is limited to a single dataset, which has unique features that may not be present in other large qualitative datasets. Another limitation is that qualitative analysis is still required on a sample of a larger dataset in order to calibrate AQUA and confirm its kappa, and is also required to generate the categories for coding in order to compare the AQUA-human ICR with the human team's ICR. Finally, AQUA is not able to match human ICRs for some categories; to analyse those categories across the entire dataset, human teams must still do the coding.

CONCLUSION

Partial automation of qualitative research studies enables researchers to conduct rigorous, rapid studies that more easily incorporate the many benefits of qualitative research. Further research is needed to determine the extent to which AQUA may be applied to other qualitative data, and the extent to which the algorithms used to augment coding may also be used to augment category development.

Acknowledgements Without the assistance of the following individuals and groups, the scope and scale of this project would not have been possible. We thank Cletis Earle, Susan Chobanoff, Neal Thomas, Leslie Parent, Sarah Bronson, Heather Stuckey-Peyrot and the rest of the Penn State Qualitative Mixed Methods Core team at Penn State.

Contributors RPL (guarantor): conceptualisation, funding acquisition, methodology, visualisation, supervision, project administration, writing-review and editing. LJVS: conceptualisation, methodology, visualisation, supervision, writing-review and editing. RF: conceptualisation, methodology, investigation, formal analysis, software, visualisation, writing-review and editing. AK: visualisation, software, writing-review and editing. XCH: visualisation, software, writing-review and editing. BLS: methodology, writing-review and editing. ELM: methodology, data curation, writing-review and editing. WAC: formal analysis, writing-review and editing. AEZ: supervision, writing-review and editing. CG: conceptualisation, methodology, investigation, formal analysis, software, supervision, visualisation, writing-review and editing.

Funding The dataset used in this work was developed with the support of the Huck Institutes of the Life Sciences (grant number 7601); the Social Science Research Institute at Penn State University (grant number 7601); and the Department of Family and Community Medicine at Penn State College of Medicine (grant number 7601-M). CG's and RF's work were supported by the Huck Institute of Life Sciences. Portions of CG's work were supported by the Defense Advanced Research Project's Agency SCORE programme (Cooperative Agreement W911NF-19-0272).

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository.

ORCID iD

Robert P Lennon http://orcid.org/0000-0003-0973-5890

REFERENCES

1 Marathe M, Toyama K. Semi-automated coding for qualitative research: a user-centered inquiry and initial prototypes. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, Montreal QC, Canada: Association for Computing Machinery, 2018:348.
2 Bryman A. Barriers to integrating quantitative and qualitative research. *J Mix Methods Res* 2007;1:8–22.
3 Wiedemann G. Opening Up to Big Data : Computer-Assisted Analysis of Textual Data in Social Sciences. *FQS* 2013;14.
4 Lewins A, Silver C. *Using software in qualitative research: a step-by-step guide*. 2nd ed. London: Thousand Oaks, 2014.
5 Guetterman TC, Chang T, DeJonckheere M, *et al*. Augmenting qualitative text analysis with natural language processing: methodological study. *J Med Internet Res* 2018;20:e231.
6 Chang T, DeJonckheere M, Vydiswaran VGV, *et al*. Accelerating mixed methods research with natural language processing of big text data. *J Mix Methods Res* 2021;15:398–412.
7 Abram MD, Mancini KT, Parker RD. Methods to integrate natural language processing into qualitative research. *Int J Qual Methods* 2020;19:160940692098460.
8 Chen H, Martin B, Daimon CM, *et al*. Effective use of latent semantic indexing and computational linguistics in biological and biomedical applications. *Front Physiol* 2013;4:8.
9 Gutu G, Dascalu M, Rebedea T. Time and semantic similarity – what is the best alternative to capture implicit links in CSCL conversations? *12th International Conference on Computer Supported Collaborative Learning (CSCL) 2017*, Philadelphia, PA: International Society of the Learning Sciences, 2017.
10 Lowe RE, Berry MW. Using non-negative tensor decomposition for unsupervised textual influence modeling. In: Berry MW, Mohamed A, Yap BW, eds. *Supervised and unsupervised learning for data science*. Cham, Switzerland: Springer International Publishing, 2020: 59–82.
11 Manning CD, Clark K, Hewitt J, *et al*. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proc Natl Acad Sci U S A* 2020;117:30046–54.

12 Griffin C, Bickel B. Unsupervised machine learning of open source Russian Twitter data reveals global scope and operational characteristics. *ArXiv* 2018.
13 Leeson W, Resnick A, Alexander D, *et al*. Natural language processing (Nlp) in qualitative public health research: a proof of concept study. *Int J Qual Methods* 2019;18:160940691988702.
14 Vollmer S, Mateen BA, Bohner G, *et al*. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ* 2020;368:l6927.
15 Sacha D, Zhang L, Sedlmair M, *et al*. Visual interaction with dimensionality reduction: a structured literature analysis. *IEEE Trans Vis Comput Graph* 2017;23:241–50.
16 Graham M, Kennedy J. A survey of multiple tree visualisation. *Inf Vis* 2010;9:235–52.
17 Schulz H-J, Hadlak S, Schumann H. The design space of implicit hierarchy visualization: a survey. *IEEE Trans Vis Comput Graph* 2011;17:393–411.
18 Wang W, Wang H, Dai G. Visualization of large hierarchical data by circle packing. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Montréal, Québec, Canada: Association for Computing Machinery, 2006:517–20.
19 CH Y, Jannasch-Pennell A, DiGangi S. Compatibility between text mining and qualitative research in the perspectives of grounded theory, content analysis, and reliability. *Qualitative Report* 2011;16:730.
20 Strickland E. How IBM Watson Overpromised and Underdelivered on AI Health Care. Institute of Electrical and Electronic Engineers Spectrum [internet], 2019. Available: https://www.mit.bme.hu/system/files/oktatas/targyak/9890/How_IBM_Watson_Overpromised_and_Underdelivered_on_AI_Health_Care_-_IEEE_Spectrum.pdf
21 Lenzner T, Neuert C, Otto W. *Cognitive Pretesting. GESIS survey guidelines*. Mannheim, Germany: GESIS – Leibniz Institute for the Social Sciences, 2016.
22 Lavrakas PJ, ed. *Encyclopedia of Survey Research Methods*. Thousand Oaks, CA: Sage Publications, Inc, 2008. htt ps://10.4135/9781412963947
23 Smith HJ, Chen J, Liu X. Language and rigour in qualitative research: problems and principles in analyzing data collected in mandarin. *BMC Med Res Methodol* 2008;8:44.
24 O'Connor C, Joffe H. Intercoder reliability in qualitative research: debates and practical guidelines. *Int J Qual Met* 2020:19.
25 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.
26 O'Connor C, Joffe H. Intercoder reliability in qualitative research: debates and practical guidelines. *Int J Qual Methods* 2020;19:1609406919899220.
27 Van Scoy L, Snyder B, Miller E. Public anxiety and distrust due to perceived politicization and media sensationalism during early COVID-19 media messaging. *J Commun Healthc* 2021.
28 Van Scoy LJ, Miller EL, Snyder B, *et al*. Knowledge, perceptions, and preferred information sources related to COVID-19 among central Pennsylvania adults early in the pandemic: a mixed methods cross-sectional survey. *Ann Fam Med* 2021;19:293–301.
29 Hsieh H-F, Shannon SE. Three approaches to qualitative content analysis. *Qual Health Res* 2005;15:1277–88.
30 Rajtmajer S, Simhachalam A, Zhao T. A dynamical systems perspective reveals coordination in Russian Twitter operations. *ArXiv* 2020.
31 Angelov D. Top2Vec: distributed representations of topics. *ArXiv* 2020.
32 Soleimani H, Miller DJ. Parsimonious topic models with salient word discovery. *IEEE Trans Knowl Data Eng* 2015;27:824–37.
33 Higham DJ, Kalna G, Kibble M. Spectral clustering and its use in bioinformatics. *J Comput Appl Math* 2007;204:25–37.
34 Andrade D, Takeda A, Fukumizu K. Robust Bayesian model selection for variable clustering with the Gaussian graphical model. *Stat Comput* 2020;30:351–76.
35 Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. *Philos Trans A Math Phys Eng Sci* 2017;374:20150202.
36 Zheng N, Xue J. *Statistical learning and pattern analysis for image and video processing advances in pattern recognition*. London: Springer, 2009.
37 Bazaraa MS, Jarvis JJ, Sherali HD. *Linear programming and network flows*. 4th ed. Hoboken, NJ: John Wiley & Sons, 2010.
38 Bottou L, Curtis FE, Nocedal J. Optimization methods for large-scale machine learning. *SIAM Rev Soc Ind Appl Math* 2018;60:223–311.
39 Belkin M, Niyogi P, Sindhwani V. *On Manifold Regularization. AISTATS 2005 - Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, 2005: 17–24. https://www.semanticscholar.org/paper/On-Manifold-Regularization-Belkin-Niyogi/b7ed5131f83783a43705db78ac5c05034659893

Supplemental Material

Table of Kappa values evaluating the inter-rater reliability or agreement between human-coders and supervised training of all 3-topic training models. Topic labels are 1: Distrust, 2: Media Messaging, 3: Trusted Sources of Information, 4: Personal Medical Concerns, 5: Family Concerns, 6: Societal Concerns, 7: Barriers to Recommendations, 8: No Worries, 9: Other Broad.

	Train Percentage			
Topics	0.2	0.4	0.6	0.8
245	0.67 (+/- 0.04)	0.69 (+/- 0.03)	0.71 (+/- 0.03)	0.72 (+/- 0.06)
145	0.67 (+/- 0.02)	0.71 (+/- 0.02)	0.70 (+/- 0.02)	0.72 (+/- 0.02)
124	0.67 (+/- 0.01)	0.70 (+/- 0.02)	0.68 (+/- 0.03)	0.70 (+/- 0.04)
157	0.65 (+/- 0.01)	0.67 (+/- 0.03)	0.67 (+/- 0.02)	0.70 (+/- 0.04)
257	0.68 (+/- 0.02)	0.70 (+/- 0.01)	0.71 (+/- 0.02)	0.70 (+/- 0.02)
345	0.64 (+/- 0.03)	0.67 (+/- 0.03)	0.68 (+/- 0.03)	0.69 (+/- 0.05)
247	0.65 (+/- 0.02)	0.68 (+/- 0.01)	0.68 (+/- 0.01)	0.69 (+/- 0.03)
235	0.62 (+/- 0.07)	0.61 (+/- 0.08)	0.61 (+/- 0.05)	0.68 (+/- 0.08)
256	0.63 (+/- 0.03)	0.66 (+/- 0.01)	0.68 (+/- 0.02)	0.68 (+/- 0.02)
135	0.66 (+/- 0.02)	0.66 (+/- 0.02)	0.67 (+/- 0.05)	0.67 (+/- 0.06)
147	0.62 (+/- 0.02)	0.66 (+/- 0.02)	0.68 (+/- 0.03)	0.67 (+/- 0.03)
156	0.61 (+/- 0.03)	0.66 (+/- 0.02)	0.67 (+/- 0.02)	0.66 (+/- 0.03)
134	0.66 (+/- 0.04)	0.64 (+/- 0.04)	0.64 (+/- 0.04)	0.65 (+/- 0.03)
234	0.58 (+/- 0.06)	0.60 (+/- 0.05)	0.61 (+/- 0.08)	0.64 (+/- 0.11)
356	0.59 (+/- 0.03)	0.61 (+/- 0.02)	0.64 (+/- 0.01)	0.64 (+/- 0.03)
357	0.60 (+/- 0.02)	0.65 (+/- 0.01)	0.65 (+/- 0.02)	0.64 (+/- 0.03)
126	0.61 (+/- 0.02)	0.63 (+/- 0.01)	0.64 (+/- 0.02)	0.64 (+/- 0.02)
347	0.63 (+/- 0.01)	0.63 (+/- 0.01)	0.64 (+/- 0.03)	0.64 (+/- 0.02)
127	0.65 (+/- 0.03)	0.61 (+/- 0.00)	0.62 (+/- 0.03)	0.62 (+/- 0.04)
236	0.56 (+/- 0.03)	0.61 (+/- 0.02)	0.59 (+/- 0.03)	0.62 (+/- 0.04)
267	0.59 (+/- 0.01)	0.62 (+/- 0.01)	0.61 (+/- 0.01)	0.62 (+/- 0.04)
136	0.59 (+/- 0.04)	0.62 (+/- 0.01)	0.62 (+/- 0.02)	0.62 (+/- 0.02)
567	0.56 (+/- 0.02)	0.57 (+/- 0.02)	0.58 (+/- 0.03)	0.61 (+/- 0.03)
167	0.57 (+/- 0.01)	0.61 (+/- 0.03)	0.60 (+/- 0.02)	0.61 (+/- 0.01)
367	0.57 (+/- 0.02)	0.60 (+/- 0.01)	0.60 (+/- 0.02)	0.60 (+/- 0.02)
146	0.54 (+/- 0.02)	0.57 (+/- 0.02)	0.59 (+/- 0.01)	0.58 (+/- 0.03)
246	0.53 (+/- 0.04)	0.58 (+/- 0.01)	0.58 (+/- 0.02)	0.58 (+/- 0.02)
346	0.52 (+/- 0.02)	0.55 (+/- 0.02)	0.56 (+/- 0.03)	0.56 (+/- 0.06)
137	0.56 (+/- 0.03)	0.57 (+/- 0.04)	0.58 (+/- 0.05)	0.55 (+/- 0.04)
123	0.51 (+/- 0.04)	0.54 (+/- 0.02)	0.52 (+/- 0.05)	0.53 (+/- 0.05)

457	0.48 (+/- 0.02)	0.52 (+/- 0.02)	0.52 (+/- 0.01)	0.53 (+/- 0.03)
467	0.50 (+/- 0.03)	0.52 (+/- 0.03)	0.54 (+/- 0.03)	0.53 (+/- 0.02)
456	0.43 (+/- 0.03)	0.49 (+/- 0.05)	0.48 (+/- 0.06)	0.50 (+/- 0.06)
237	0.50 (+/- 0.02)	0.49 (+/- 0.05)	0.48 (+/- 0.05)	0.47 (+/- 0.05)