Multimodal machine learning in precision health: A scoping review

Adrienne Kline1, Hanyin Wang1, Yikuan Li1, Saya Dennis1, Meghan Hutch1, Zhenxing Xu2, Fei Wang2, Feixiong Cheng3 and Yuan Luo1✉

Machine learning is frequently being leveraged to tackle problems in the health sector including utilization for clinical decision-support. Its use has historically been focused on single modal data. Attempts to improve prediction and mimic the multimodal nature of clinical expert decision-making has been met in the biomedical field of machine learning by fusing disparate data. This review was conducted to summarize the current studies in this field and identify topics ripe for future research. We conducted this review in accordance with the PRISMA extension for Scoping Reviews to characterize multi-modal data fusion in health. Search strings were established and used in databases: PubMed, Google Scholar, and IEEExplore from 2011 to 2021. A final set of 128 articles were included in the analysis. The most common health areas utilizing multi-modal methods were neurology and oncology. Early fusion was the most commonly data merging strategy. Notably, there was an improvement in predictive performance when using data fusion. Lacking from the papers were clear clinical deployment strategies, FDA-approval, and analysis of how using multimodal approaches from diverse sub-populations may improve biases and healthcare disparities. These findings provide a summary on multimodal data fusion as applied to health diagnosis/prognosis problems. Few papers compared the outputs of a multimodal approach with a unimodal prediction. However, those that did achieved an average increase of 6.4% in predictive accuracy. Multi-modal machine learning, while more robust in its estimations over unimodal methods, has drawbacks in its scalability and the time-consuming nature of information concatenation.

npj Digital Medicine (2022)5:171 ; https://doi.org/10.1038/s41746-022-00712-8

INTRODUCTION

Clinical decision support has long been an aim for those implementing algorithms and machine learning in the health sphere1–3. Examples of algorithmic decision supports utilize lab test values, imaging protocols or clinical (physical exam scores) hallmarks4,5. Some health diagnoses can be made on a single lab value or a single threshold, such as in diabetes in older adults6. Other diagnoses are based on a constellation of the signs, symptoms, lab values and/or supportive imaging and are referred to as a clinical diagnosis. Oftentimes these clinical diagnoses are based on additive scoring systems that requires an admixture of positive and negative hallmarks prior to confirmatory labeling.

The modus operandi of a clinical diagnosis may fail to consider the relative weighting of these disparate data inputs and potentially non-linear relationships highlighting the limitations of human decision-making capacity. The strength of algorithmic decision-making support is that it can be used to offload such tasks, ideally yielding a more successful result. This is the premise of precision medicine. Precision medicine/health aims to create a medical model that customizes healthcare (decisions, treatments, practices etc.) that are tailored to either an individual or patient phenotype7. This includes tracking patients’ health trajectories longitudinally8, oftentimes incorporating genetics/epigenetics9,10 and mathematical modeling11 where diagnoses and treatments incorporate this unique information12. Contrast this with a one-drug-fits-all model, where there is a single treatment per disorder. Figure 1 illustrates the flow of information from hospitals/care centers that generate disparate data. It is through computational modeling and information fusion that outcomes of interest such as drug and treatment targets ultimately facilitate better decision making at the patient level in those care centers. This phenomenon has sparked an interest in fusion studies using health care data.

Undertakings to characterize this literature have been performed by Huang et al.13, who performed a systematic review of deep learning fusion of imaging and EHR data in health. However, it was limited to EHR and imaging data and deep learning applications. A follow-up review article included a commentary on omics and imaging data fusion14. The purpose of this study is to highlight the current scope of this research domain, summarize and offer suggestions to advance the field. The current study is more inclusive in the breadth of the types of machine learning protocols used and attempts to encompass all current modalities (information types/sources).

Data fusion is underpinned by information theory and is the mechanism by which disparate data sources are merged to create an information state based on the sources’ complementarity15,16 (Box 1). The expectation in machine learning is that data fusion efforts will result in an improvement in predictive power17,18 and therefore provide more reliable results in potentially low validity settings19. Data fusion touts the advantage that the results of modeling become inherently more robust by relying on a multitude of informational factors rather than a single type. However, the methodology of combinatorial information has drawbacks; it adds complexity to specifying the model and reduces the interpretability of results18,20.

Data from different sources and file formats are rarely uniform, and this is especially the case with clinical data21. For example, data sets can have different naming conventions, units of

1Department of Preventive Medicine, Northwestern University, Chicago 60201 IL, USA. 2Department of Population Health Sciences, Cornell University, New York 10065 NY, USA. 3Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland 44195 OH, USA. ✉email: yuan.luo@northwestern.edu

Published in partnership with Seoul National University Bundang Hospital
measure, or represent different local population biases. Care must be taken to search and correct for systematic differences between datasets and assess their degree of inter-operability. For example, Colubri et al. aggregated computed tomography (CT) and PCR lab values, by performing an intra-site normalization. This ensured that the values were comparable across sites. In doing so they discarded several potentially informative clinical variables since they were not all available in all datasets.

A balance is required to allow information that is similar to work together (harmonization) and retain data purity (information correspondence). Successful fusion uses data harmonization techniques that assure both in the quality control of the integration process. Clinical data harmonization requires multidisciplinary research among medicine, biology, and computer science. The clinical area of heart failure with preserved ejection fraction (HFpEF) saw novel applications of multiple tensor factorization formulations to integrate the deep phenotypic and trans-omic information, and this extends to other areas of precision medicine. To increase the portability of EHR-based phenotype algorithms, the Electronic Medical Records and Genomics (eMERGE) network has adopted common data models (CDMs) and standardized design patterns of the phenotype algorithm logic to integrate EHR data with genomic data and enable generalizability and scalability.

There are three main types of data fusion that are used in machine learning: early (data-level), intermediate (joint), and late (decision-level). In the case of early fusion, multiple data sources are converted to the same information space. This often results in vectorization or numerical conversion from an alternative state, such as that performed by Chen et al. via vectorized pathology reports. Medical images possess characteristics that can undergo numerical conversion based on area, volume, and/or structural calculations. These are then concatenated with additional measurements from structured data sources and fed into an individual classifier. Canonical correlation analysis, non-negative matrix factorization, Independent Component Analysis (ICA)

---

**Box 1 Terms and Concepts**

Multimodal machine learning: the area of machine learning concerned with bringing together disparate data sources to capitalize on the unique and complementary information in an algorithmic framework.

Data harmonization: using machine learning to unify different data sources to improve its quality and utilization.

Multiview machine learning: another term for multimodal machine learning.

Data fusion: the specific methodologies undertaken to perform data integration for multimodal/multiview machine learning; they come in three broad categories: early, intermediate/joint, and late.

---

**Fig. 1 Multimodal precision health; the flow of information.** Information moves in a cyclical pattern from health centers to information commons, where it can be transformed and algorithmic modeling performed. These algorithms provide insight into many different health outcomes such as clinical trials, phenotyping, drug discovery, etc. These insights should return to health centers and practitioners to provide the most efficient, evidence-based medicine possible.
and numerical feature conversion methodologies exist as common options to transform all data into the same feature space. Intermediate data fusion occurs as a stepwise set of models and offers the greatest latitude in model architecture. For example, a 3-stage deep neural learning and fusion model was proposed by Zhou et al. Stage 1 consists of feature selection by a soft-max classifier for independent modalities. Stages 2 and 3 constitute combining these selected features, establishing a further refined set of features, and feeding these into a Cox-net to perform joint latent feature representation for Alzheimer’s diagnosis. In contrast to early fusion, intermediate fusion combines the features that distinguish each type of data to produce a new representation that is more expressive than the separate representations from which it arose.

In late fusion, typically multiple models are trained where each model corresponds to an incoming data source. This is akin to ensemble learning, which offers better performance over individual models. Ensemble methods use multiple learning algorithms (typically applied to the same dataset) to obtain better predictive performance than could be obtained from any of the constituent learning algorithm alone. However, multimodal machine learning ensemble here can refer to ensemble learning within a data type or across data types. These take symbolic representations as sources and combine them to obtain a more accurate decision. Bayesian’s methods are typically employed at this level to support a voting process between the set of models into a global decision. Within late fusion there has been headway made to perform multitask deep learning. A schematic for the 3 subtypes of data fusion is presented in Fig. 2. Attributes in the fusion techniques are shown in Table 1.

**RESULTS**

**Topic Modeling**

The topic modeling displayed in Fig. 3 showcases the category, specific health ailments under investigation, and the modality type for the studies included. These were subsequently mapped to the category of the combination of information that were merged to create models for prediction/classification/clustering (Table 2). This plot should serve as a resource to fellow researchers to identify areas that are less frequent, such as dermatology, hematology, medication/drug issues such as alcohol use disorder that may offer new research horizons. Figure 4 identifies coding platforms, publishing trend and location over time, author locations and patient cohorts of the papers included in this review.

| Attribute                        | Early | Intermediate/joint | Late/decision |
|----------------------------------|-------|--------------------|---------------|
| Scalable                         | No    | Yes                | Yes           |
| Multiple models needed           | No    | Yes                | Yes           |
| Improved accuracy                | Yes   | Yes                | Yes           |
| Voting of multiple models        | No    | Yes                | Yes           |
| Interaction effects across sources | Yes  | Yes                | No            |
| Implemented in health            | Yes   | Yes                | Yes           |

*Fig. 2  Early, intermediate, and late fusion; flow of information from information commons to model structure to outcomes. Information fusion can occur in a myriad of ways. In machine learning, early, intermediate, and late fusion is typified by if all the information flows into a single model (early), a step-wise fashion where outputs from one model become inputs for another (intermediate), and lastly, where all unique data types undergo separate modelling after which ensembling and/or voting occurs (late).*
Of the models used in the papers, 126/128 explicitly reported performing a validation procedure of them. The most common validation processes performed were N-fold cross validation (55), train test split (51), leave one out cross validation (10), and external dataset (10). A cornucopia of machine learning techniques and methods were used within and across articles in this review. They have been summarized in Table 2, noting in which fusion umbrella subtype they were implemented.

Early fusion

Most papers were published using early fusion. Of those, most were published using medical imaging and EHR data (51,52,56). Nearly all these papers performed numericalization of image features in essence converting them to structured data prior to processing, however, two performed matrix factorization (33,36). A combination of EHR and text data was noted in 15 papers (31,54,69,72,79–81,91–93). Meng et al. created a Bidirectional Representation Learning model used latent Dirichlet allocation (LDA) on clinical notes (112). Cohen et al. used unigrams and bigrams in conjunction with medication usage (54). Zeng et al. used concept identifiers from text as input features (31). Ten papers used early fusion with using EHR and genomic data (12,50,51,55,65,83,89,108). Doan et al. concatenated components derived from images with polygenic risk scores (33). Lin et al. et al. created aggregated scores from MRI, cerebral spinal fluid, and genetic information and brought them together into a single cohesive extreme learning machine to predict mild cognitive impairment (55). Tremblay et al. used a multivariate adaptive regression spline (MARS) after normalizing, removing highly correlated features (89). Ten papers performed fusion using imaging and genomic data (33,52,70,76–78,82,84,97,110). Three of these generated correlation matrices as features by vectorizing imaging parameters and correlating them with single nucleotide polymorphisms (SNPs) prior to feeding into the model (33,70,77). Three papers in this category used EHR and time series (58,74,101). Both Hernandez and Cannière et al. implemented their methods for purposes of cardiac rehabilitation and harnessed the power of support vector machines (SVMs). However, Hernandez preserved time series information by assembling ECG data into tensors that preserve the structural and temporal relationships inherent in the feature space (74), while Cannière performed dimensionality reduction of

---

**Fig. 3** **Topic and Modality Modeling.** Neurology, and in particular, Alzheimer’s disease investigations accounted for the most papers published on this topic (n = 22). With the onset of the COVID-19 pandemic, several primary research articles were dedicated to this topic, which can be arrived at through the respiratory or infectious disease hierarchies. All papers noted in this review used either two or three disparate data sources when fusing their data, and specifically that of imaging and EHR (n = 52), was the most prevalent.
the time series information using t-SNE plots. Two papers comprised early fusion using imaging and time series. There were two papers that leveraged EHR and genomic information. Luo et al. implemented hybrid non-negative matrix factorization (HNMF) to find coherence between phenotypes and genotypes in those suffering from hypertension. One paper leveraged early fusion using imaging and text data and another used EHR, Genomics, Transcriptomics, and Insurance Claims.

**Intermediate fusion** had the second highest number of papers published. 14 used imaging and EHR data. Zihni et al. merged the output from a Multilayer Perceptron (MLP) for modeling clinical data and convolutional neural network (CNN) for modeling imaging data into a single full connected final layer to predict stroke. A very similar approach...

---

**Table 2.** Fusion and machine learning methods included in this review.

| Fusion type         | Machine learning models/techniques implemented                                                                 | References                                                                 |
|---------------------|---------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------|
| Early               | Support Vector Machine, Random Forest, Decision Trees, Gaussian model, Bayesian models: Bayesian network, Naïve Bayes, Regression: Logistic, Ridge, Cox, LASSO, MARS, Multitask learning, iMSF, Boosted Models, Adaboost, Deep Neural Networks: DNN, RNN, CNN, DUN, AutoEncoders, Natural Language Processing, GPT, BERT, Clustering: K-means, hierarchical, KNN, Graph models, Ensemble learning, Artificial Neural Networks, Markov Model, Transfer learning, Multitask learning. | 31–34,36,50,51,53–81, 31,33,34,36,67–71,73–76,78–80,82–91, 36,55,92, 31,54,71,73–75,80,93, 31,51,67–69,77,79,80,87,89,90,94–99, 62, 100, 68,71–75,79,90, 48,57,61,80,81,91,97,101–107, 81,102,105, 65,73,75,76,100,108, 107–109, 65,88, 57,64,69,73,78,81,88,91,101,104,107,110,111, 108, 48,62,97,105,112, 62, 41,43,47,59,113–122, 37,59,113–115,117,120,123–126, 41,40,59,115–117,120,122,127,128, 37,41–47,59,115–117,121–135, 116,118,126, 122,126,134,136, 43–45,49,116,117,121,126,135, 59, 120, 46, 47,137, 136, 138–143, 130,140–147, 141,143,147,148, 39,130,140,142,144–146,149–153, 154, 43,47,59,113,120, 143,147,153, 39,153,155, 145,153,155, 140–143,147,155,156, 157, 144,145,157, 158, 39,153,155, 145,153,155, 140–143,147,155,156, 157, 144,145,157, 158, 39,153,155, 145,153,155, 140–143,147,155,156, 157, 144,145,157, 158. |
| Intermediate        | Regression: logistic, multivariate, support vector, LASSO, Support Vector Machine, Decision Trees, Random Forest, Deep Learning: CNN, DNN, RNN, AutoEncoders, Boosted Learning: gradient boosting, XGBoost, Adaboost, Natural Language Processing, Artificial Neural Network, Naïve Bayes, Ensemble Learning, Transfer Learning, Multitask learning, Graph Learning. | 41,43,47,59,113–122, 37,59,113–115,117,120,123–126, 41,40,59,115–117,120,122,127,128, 37,41–47,59,115–117,121–135, 116,118,126, 122,126,134,136, 43–45,49,116,117,121,126,135, 59, 120, 46, 47,137, 136, 138–143, 130,140–147, 141,143,147,148, 39,130,140,142,144–146,149–153, 154, 43,47,59,113,120, 143,147,153, 39,153,155, 145,153,155, 140–143,147,155,156, 157, 144,145,157, 158, 39,153,155, 145,153,155, 140–143,147,155,156, 157, 144,145,157, 158. |
| Late                | Support Vector Machine, Random Forest, Decision Trees, Regression: LASSO, Logistic, Multivariate, GAM, Deep Learning: CNN, DNN, RNN, AutoEncoders, Ensemble learning, Multitask learning, XGBoost, Artificial neural network, Natural Language Processing, Regression: GLM, Logistic, Graph network, Clustering: KNN, graphs, Mixed Random Forest, Artificial Neural Network, ElasticNet. | 138–143, 130,140–147, 141,143,147,148, 39,130,140,142,144–146,149–153, 154, 43,47,59,113,120, 143,147,153, 39,153,155, 145,153,155, 140–143,147,155,156, 157, 144,145,157, 158, 39,153,155, 145,153,155, 140–143,147,155,156, 157, 144,145,157, 158. |

CNN convolutional neural network, DUN deep unified networks, MARS Multivariate adaptive regression splines, MLP multilayer perceptron, GLM Generalized Linear Model, DNN deep neural network, GAM generalized additive model, iMSF incomplete Multi-Source Feature, RNN Recurrent Neural Network.
was taken by Tang et al. who used three-dimensional CNNs and merged the layers in the last layer. EHR and text data were fused together in 11 papers. Of these, six used long term short term (LSTM) networks, CNNs, or knowledge-guided CNNs in their fusion of EHR and clinical notes. Chowdhury et al. used graph neural networks and autoencoders to learn meta-embeddings from structured lab test results and clinical notes. Pivovarov et al. learned probabilistic phenotypes from clinical notes and medication/lab orders (EHR) data. Two models each employing LDA where data type was treated as a bag of elements and to bring coherence between the two models to identify unique phenotypes. Ye et al. and Shin et al. used concept identifiers via NLP and bag-of-words techniques, respectively, prior to testing a multitude of secondary models. In general, clinical notes can provide complementary information to structured EHR data, where natural language processing (NLP) is often needed to extract such information. A few studies were published using imaging and genomics. Here radiogenomics were used to diagnose attention-deficit/hyperactivity disorder (ADHD), glioblastoma survival, and dementia respectively. Polygenic risk scores were combined with MRI by Yoo et al. who used an ensemble of random forests for ADHD diagnosis. Zhou et al. fused SNPs information together with MRI and positron emission tomography (PET) for dementia diagnosis by learning latent representations (i.e., high-level features) for each modality independently. Subsequently learning joint latent feature representations for each pair of modality combination and then learning the diagnostic labels by fusing the learned joint latent feature representations from the second stage was carried out. Wijethilake used MRI and gene expression profiling, performing recursive feature elimination prior to merging into multiple models SVM, linear regression, and artificial neural network (ANN). The linear
regression model outperformed the other two merged models and any single modality\textsuperscript{177}. Wang et al. and Zhang et al. showcased their work in merging imaging and text information\textsuperscript{146,147}. Both used LSTM for language modeling a CNN to generate embeddings that were joined together in a dual-attention model. This is achieved by computing a context vector with attended information preserved for each modality resulting in joint learning. Seldom were articles published using: Imaging/EHR/Text\textsuperscript{113}, Genomic/Text\textsuperscript{69}, Imaging/Time series\textsuperscript{127}, Imaging/Text/Time series\textsuperscript{47}, Imaging/EHR/Genomic\textsuperscript{128}, Imaging/EHR/Time series\textsuperscript{124}, EHR/Genomic\textsuperscript{128}, EHR/Text/Time series\textsuperscript{42}.

**Late fusion**

A much smaller number ($n = 20$) of papers used late fusion. Seven of those used imaging and EHR data types\textsuperscript{138,139,144,150,151,154,164}. Both Xiong et al. and Yin et al. fed outputs into a CNN to provide a final weighting and decision\textsuperscript{150,151}. Three papers were published using a trimodal approach: imaging, EHR, and genomic\textsuperscript{130,147,148}. Xu et al. and Faris et al. published papers using EHR and text data\textsuperscript{146,155}. Faris et al. processed clinical notes using TF-IDF, hashing vectorizer and document embeddings in conjunction with binarized clinical data\textsuperscript{155}. Logistic Regression (LR), Random Forest (RF), Stochastic Gradient Descent Classifier (SGD Classifier), and a Multilayer Perceptron (MLP) were applied to both sets of data independently and final outputs of the two models were combined using different schemes: ranking, summation, and multiplication. Two articles were published using imaging and time series\textsuperscript{149,152} both of which employed CNNs, one in video information of neonates\textsuperscript{149} and the other in chest x-rays\textsuperscript{152}. However, they differed in their processing of the time series data. Salekin used a bidirectional CNN and Nishimori used a one-dimensional CNN. Far fewer papers were published using Imaging/EHR/Text\textsuperscript{133}, EHR/Genomic/Text\textsuperscript{145}, imaging/EHR, time series\textsuperscript{141}, Imaging/Genomic\textsuperscript{136}, EHR/Genomic\textsuperscript{140}, and Imaging/Text\textsuperscript{99}.

**Mixed fusion**

Two papers performed multiple data fusion architectures\textsuperscript{158,159}. Huang et al. created seven different fusion architectures. These included, early, joint, and late fusion. The architecture that performed the best was the late elastic average fusion for the diagnosis of pulmonary embolism using computed tomography and EHR data\textsuperscript{159}. Their Late Elastic Average Fusion leveraged an ElasticNet (linear regression with combined L1 and L2 priors that act as regularizers) for EHR variables. El-Sappagh et al. performed early and late fusion to create an interpretable Alzheimer’s diagnosis and progression detection model\textsuperscript{158}. Their best performing model was one that implemented instance-based explanations of the random forest classifier by using the SHapley Additive exPlanations (SHAP) feature attribution. Despite using clinical, genomic, and imaging data, the most influential feature was found to be the Mini-Mental State Examination.

**Clinical relevance**

Data fusion may help address sex representation and increase population diversity issues (including minority populations) in health modeling by creating a more representative dataset if one datatype contained more of one sex and another datatype contained more of the other. This reciprocal compensation ability of employing various data sets would also hold true for racial or ethnic diversities.

Less than half (37.6\%) of the papers were published in a journal intended for a clinical audience. None of the papers included in the final cohort of studies had created tools for clinical use that had FDA approval. Based on the rising number of papers in this field there is a growing and global need and interest to characterize these findings.

**DISCUSSION**

Returning to our research questions, we outlined from the inception of this work, we arrive at Table 3.

### Table 3. Research questions as outlined in Methods.

| RQ1 | The literature published in this area as displayed and characterized in the Results’ section is one that is of growing and global interest. |
| RQ2 | The analysis techniques are varied and currently do not show a standard machine learning method in the field. This is likely linked to it being a relatively new and emerging field. |
| RQ3 | Health contexts predominantly impacted by this include Neurology and Cancer. |

Published in partnership with Seoul National University Bundang Hospital

A. Kline et al.  npj Digital Medicine (2022) 171
Many issues were raised in the papers included in this review. The most common reported limitations were cohorts from a single site, small sample sizes, retrospective data, imbalanced samples, handling of missing data, feature engineering, controlling for confounding factors, and interpretation of the models employed. Samples were most often built from a single hospital or academic medical center. Small sample sizes often lead to poor model fitting and generalizability. The median number of unique patients reported across the studies was 658 with a standard deviation of 42,600. This suggests that while some studies were able to leverage large and multi-center cohorts, a great many were not able to do so.

Seldom were machine learning investigations on prospective data, an issue endemic in the field. Sample imbalances were often ignored, which results in biased models and misleading performance metrics. Missing data were usually ignored by dropping data or imputing, if not dealt with appropriately can skew the results. More studies need to discuss frequencies and types of missing data. Comparison of different imputation methods on the final results should be part of the reporting process. When performing statistical analysis, researchers usually ignored possible confounding factors such as age or gender. Doing so may have major effects on the impact of results. Such possible confounding effects should either be taken into consideration by the model or adjusted for first, prior to reporting model results. Reasonable interpretations of the model and outputs must be presented so that clinicians find the results credible and then use them to provide guidance for treatments. However, most authors did not take the time to interpret the models for clinical audiences. Additionally, how the results may function as a clinical decision support tool. Different types of models warrant different explanations. These limitations are highlighted where they occur in the data processing and modeling building pipeline in Fig. 5.

To expedite and facilitate this field, we have outlined several gaps for future research in this field. These are listed in Fig. 5 and explored. Medication/drug topics present an underrepresented area, with only two papers being published in this field. Awareness of drug interaction effects is a difficult and growing issue, particularly in geriatrics, which gave rise to Beer’s criteria. Performing multimodal machine learning may offer an earlier detection of adverse events associated with medication misuse that is a result of iatrogenic error, non-compliance, or addiction. Similar justifications as outlined above could be applied to other areas seen as ‘under saturated’ such as hematology with only one paper and nephrology having just three.

Augmenting clinical decision-making with ML to improve clinical research and outcomes offers positive impacts that have
economic, ethical, and moral ramifications, as it can reduce suffering and save human lives. Multiple studies have now pointed out that if the data an ML model is trained on is biased this often yields bias in the predictions. Ensuring multisite, representative data will limit model biases. We also advocate for the creation of open access pipelines/libraries to speed up data conversion to make the technology more widely available. Improving accuracy at the expense of complex and time-consuming data transformations may mean the predictive power gained from a multimodal approach is offset by this front-end bottleneck, meaning predictions are no longer temporally relevant or useful.

While incorporating disparate data does lend itself to seemingly better predictions, as knowledge around certain diseases accumulates, data fusion in healthcare is an evolving target that warrants proactively adapting to the dynamic landscape. There is no single ML model with ubiquitous applicability. For example, it has been shown in protein-protein interactions that utilization of the XGBoost ensemble algorithm reduces noisy features, maintains the significant raw features, and prevents overfitting. Similarly, LightGBM has the advantages of faster training speed, higher efficiency, lower memory usage, better accuracy, and has been consistently outperforming other models. Graph neural networks can synthesize new connections leading to drug discovery/targets.

In the same vein, models that permit interpretability should always be considered. For example, the Perotte et al. model was not compared with conventional simpler machine learning classifiers, and collective matrix factorization becomes inherently difficult to interpret. Contrast this with the work of Fracarro et al. whose study of macular degeneration noted their white box performed as well as black box methods implementations.

As this field and the datasets associated mature there is work needed to address the tenets of data management: Findability, Accessibility, Interoperability, and Reuse of digital datasets (FAIR). This entails having metadata that are unique/de-identified and searchable, with open or federated access points (Findability/Accessibility), data that are shared broadly (Interoperable), and finally data that contain accurate and relevant attributes under a clear data usage agreement/license (Reusable). It is imperative there exist a clear definition of outcomes, assessment of biases and interpretability/ transparency of results, and limitations inherent in its predictions.

Of crucial importance for uptake is that predictions be patient-specific and actionable at a granular level. For example, a 30-day readmission prediction algorithm if implemented, may inform resource management and prompt additional research that may decrease the number of patients re-admitted. Linden et al. developed Deep personalized LOnitudinal convolutional Risk model (DeepLOR) capable of creating predictions that can be interpreted on the level of individual patients. Leveraging both and clinical and empirically driven information to create meaningful and usable recommendations may improve clinician/end-user understanding by relating to existing frameworks. Resources such as CRISP-ML provide a framework for moving use cases into more practical applications, while efforts to vie for Food and drug administration (FDA) approvals as a tool for use are encouraged to increased adoption.

Deployment of models with user interfaces annotating limitations inherent on those predictions will allow clinical decision makers to interface and implement change accordingly. Follow-through on the aforementioned tasks will push individual fields to create recommendations for subsequent real-world implementations that are relevant, actionable, and transcend regional/subpopulation differences. Limitations of this scoping review include that it is not a systematic review. Therefore, it is possible that some titles that should have been included were missed. As the primary purpose of this study was to perform scientific paper profiling on multimodal machine learning in health, a critical appraisal of individual methodological quality of the included studies was not performed. However, commentary is provided on the methodological limitations that could have affected their results and impacted their claims. This review offers comprehensive meta-data and evaluation across health domains, immaterial to the type of machine learning or the data used. This work serves as both a summary and steppingstone for future research in this field. Data fusion in health is a growing field of global interest. The topic areas of health that have high frequency relative to others were neurology and cancer, which serve to highlight opportunities for further exploration in understudied topics (hematology, dermatology). Unimodal machine learning is inherently in contrast to current routine clinical practice in which imaging, clinical or genomic data are interpreted in unison to inform accurate diagnosis and warrants further work for ease of use and implementation. Overall, it appears justified to claim that multimodal data fusion increases predictive performance over unimodal approaches (6.4% mean improvement in AUC) and is warranted where applicable. Multimodal machine learning may be a tool leveraged in precision medicine to further subgroup patients' and their unique health fingerprint. Furthermore, as no papers in our review sought FDA approval, we advocate for more efforts into model translation and explore necessities that facilitate that end.

A dashboard resource published in conjunction with this review article is available at: https://multimodal-ml-health.herokuapp.com/. This dashboard was created as an interactive infographic-based display of the major findings presented in this paper. To foster future work, a drop-down menu was created to help researchers filter the underlying data file of titles based on the specific overarching health topic by selection. This will facilitate the location of relevant papers.

**METHODS**

**Search strategy and selection criteria**

Inclusion requirements were: (a) original research article; (b) published within the last 10 years (encompassing years 2011–2021); (c) published in English; and (d) on the topic of multi-modal or multi-view using machine learning in health for diagnostic or prognostic applications. Multi-modal’ or multi-view’ for our context means the multiple data sources were not of the same type. For example, while a paper using CT and MRI may be considered multi-modal imaging; however, under our criteria it would be considered uni-modal (i.e., only included imaging). Exclusions for the purposes of this review were: (a) scientific articles not published in English; (b) commentaries or editorials; or (c) other review articles. Papers were also excluded if the data were not human-derived. We also excluded papers where the fusion already occurred at the data generation stage, such as spatial transcriptomics producing integrated tissue imaging and transcriptomics data. All papers underwent a 2-person verification for inclusion in the manuscript.

Search strings were established via literature searches and domain expertise. Additional keywords were identified based on key word co-occurrence matrices established from the abstracts of the previously included articles. Figure 6a displays the search strings, where an individual string would include one keyword from each column, this was performed for all combinations of search strings. An overview of the inclusion/exclusion process is noted in Fig. 6b and follows the standard set by PRISMA extension for scoping reviews.

Published in partnership with Seoul National University Bundang Hospital  npj Digital Medicine (2022) 171
were extracted by reviewing the primary paper. The significance and outcomes measures were all recorded for each publication (yes/no), the statistical tests run, the nature of the machine learning algorithms used. Whether validation (c) data merging technique (early, intermediate, late), and (d) interface(s) used in data processing/analysis, (b) machine learning type, (a) heterogeneous data + multi-modal machine learning. (article, disease in question. It arises from the topics of Cancer and Respiratory according to our classification. Health topic classification was overseen and reviewed by a medical doctor to ensure accuracy. As multiple health topics often encompassed a single health disease addressed in each paper, several papers are counted twice. This is true when being mapped from the right side of the Sankey plot to the specific health disease in the middle.

We recorded and extracted the number of different modalities and the divisions (i.e., text/image vs EHR/genomic/time series) used. The objective of each paper was extracted in a 1–2 sentence summary along with the keyword (if available). Patient characterization in the studies was performed by ascertaining the number of unique patients in the cohort and patient sex (i.e., Men/women/both or not mentioned).

Computational information extracted included: (a) the coding interface(s) used in data processing/analysis, (b) machine learning type, (c) data merging technique (early, intermediate, late), and (d) types of machine learning algorithms used. Whether validation was performed (yes/no), the statistical tests run, the nature of the validation, and outcomes measures were all recorded for each paper. The significance, impact, and limitations of each paper were extracted by reviewing the primary findings and limitations as noted in the papers.

### Data extracted

Information garnered from the articles included title, year published, FDA approval of the tool, whether published in a clinical journal, author affiliations, number of authors, locations (continents), and abstract. Health topic(s) addressed were extracted, as well as the broader medical topic(s) that encompass the disease. For example, lung cancer would be the specific disease in question. It arises from the topics of Cancer and Respiratory according to our classification. Health topic classification was overseen and reviewed by a medical doctor to ensure accuracy. As multiple health topics often encompassed a single health disease addressed in each paper, several papers are counted twice. This is true when being mapped from the right side of the Sankey plot to the specific health disease in the middle.

We recorded and extracted the number of different modalities and the divisions (i.e., text/image vs EHR/genomic/time series) used. The objective of each paper was extracted in a 1–2 sentence summary along with the keyword (if available). Patient characterization in the studies was performed by ascertaining the number of unique patients in the cohort and patient sex (i.e., Men/women/both or not mentioned).

Computational information extracted included: (a) the coding interface(s) used in data processing/analysis, (b) machine learning type, (c) data merging technique (early, intermediate, late), and (d) types of machine learning algorithms used. Whether validation was performed (yes/no), the statistical tests run, the nature of the validation, and outcomes measures were all recorded for each paper. The significance, impact, and limitations of each paper were extracted by reviewing the primary findings and limitations as noted in the papers.

### Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

---

**Data extracted**

Information garnered from the articles included title, year published, FDA approval of the tool, whether published in a clinical journal, author affiliations, number of authors, locations (continents), and abstract. Health topic(s) addressed were extracted, as well as the broader medical topic(s) that encompass the disease. For example, lung cancer would be the specific disease in question. It arises from the topics of Cancer and Respiratory according to our classification. Health topic classification was overseen and reviewed by a medical doctor to ensure accuracy. As multiple health topics often encompassed a single health disease addressed in each paper, several papers are counted twice. This is true when being mapped from the right side of the Sankey plot to the specific health disease in the middle.

We recorded and extracted the number of different modalities and the divisions (i.e., text/image vs EHR/genomic/time series) used. The objective of each paper was extracted in a 1–2 sentence summary along with the keyword (if available). Patient characterization in the studies was performed by ascertaining the number of unique patients in the cohort and patient sex (i.e., Men/women/both or not mentioned).

Computational information extracted included: (a) the coding interface(s) used in data processing/analysis, (b) machine learning type, (c) data merging technique (early, intermediate, late), and (d) types of machine learning algorithms used. Whether validation was performed (yes/no), the statistical tests run, the nature of the validation, and outcomes measures were all recorded for each paper. The significance, impact, and limitations of each paper were extracted by reviewing the primary findings and limitations as noted in the papers.
68. Paolo, F. et al. Combining macula clinical signs and patient characteristics for age-related macular degeneration diagnosis: A machine learning approach. Retina. BMC Ophthalmol. 15, 10 (2015).
69. Benjamin, D. W. et al. Early identification of epilepsy surgery candidates: A multicenter, machine learning study. Acta Neurol. Scand. 144, 41–50 (2021).
70. Xia An, B., Wenyuan, Z., Lou, L. & Zhaoxu, X. Detecting risk gene and pathogenic brain region in EMG using a novel GFRP algorithm based on brain imaging and genetic data. IEEE J. Biomed. Health Inform. 25, 3019–3028 (2021).
71. Prashanth, R., Sumanta Dutta, R., Pravat, K. M. & Shantanu, G. High-accuracy detection of early Parkinson’s disease through multimodal features and machine learning. Int. J. Med. Inf. 90, 13–21 (2016).
72. Ali, A.-M. et al. Machine learning for localizing epileptogenic zone in the temporal lobe: Quantifying the value of multimodal clinical-semiology and imaging concordance. Front. Digital Health 3, 559103 (2021).
73. Baiying, L. et al. Assessment of liver fibrosis in chronic hepatitis B via multimodal data. Neurocomputing 253, 169–176 (2017).
74. Larry, H. et al. Multimodal tensor-based method for integrative and continuous livers disease using machine learning. J. Healthcare Inform. Res. 26, 3541–3551 (2019).
75. Wei Liang, T., Chee Kong, C., Sim Heng, O. & Alvin Choong Meng, N. Ensemble-based regression analysis of multimodal medical data for osteopenia diagnosis. IEEE J. Biomed. Health Inform. 4, 283–299 (2019).
76. Eleftherios, T., Ioannis, S., Apostolos, H. K. & Kostas, M. Deep radiotranscriptomics for multimodal representation learning across medical images and reports. BMC Med. Inform. Decis. Mak. 21, 101977 (2020).
77. Kline et al. Multimodal clinical-skeletal data for predicting chronic obstructive pulmonary disease through deep learning. Expert Syst. Appl. 58, 811–819 (2013).
78. Sebastien, T., Yasser, I.-M., Jose Maria, M.-P., Alan, C. E. & Louis De, B. Defining a multimodal signature of remote sports concussions. Eur. J. Neurosci. 46, 1956–1967 (2017).
79. Gianluca, B. et al. Multimodal predictive modeling of endovascular treatment outcome for acute ischemic stroke using machine-learning. Stroke 51, 1285–1296 (2020).
80. Yiming, X. et al. Explainable dynamic multimodal variational autoencoder for the prediction of patients with suspected central precocious puberty. IEEE J. Biomed. Health Inform. 26, 1362–1373 (2021).
81. Alan, D. K. et al. Mixture model framework for traumatic brain injury prognosis using heterogeneous clinical and outcome data. IEEE J. Biomed. Health Inform. 25, 1321–1329 (2021).
82. Casper, R. et al. Preoperative risk stratification in endometrial cancer (ENDORISK) by a Bayesian network model: A development and validation study. PLoS Med. 17, e1003111 (2020).
83. Tommaso, G. et al. SARS-COV-2 comorbidity network and outcome in hospitalized patients in Crema, Italy. PLoS One 16, e0248498 (2021).
84. Huan, Q. et al. Machine-learning radiomics to predict early recurrence in perihilar cholangiocarcinoma after curative resection. Liver Int. 41, 837–850 (2021).
85. Ramon, C. et al. Alzheimer’s disease risk assessment using large-scale machine learning methods. PLoS One 8, e77949 (2013).
86. Aleksiev, T. et al. Multimodal machine learning-based knee osteoarthritis progression prediction from plain radiographs and clinical data. Sci. Rep. 9, 20038 (2019).
87. Michael, J. D. et al. Development and validation of a novel automated Gleason grade and morphologic profile that define a highly predictive prostate cancer progression algorithm-based test. Prostate Cancer Prostatic Dis. 21, 594–603 (2018).
88. Perotte, A., Ranganath, R., Hirsch, J. S., Blei, D. & Elhadad, N. Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. J. Am. Med. Inf. Assoc. 22, 872–880 (2015).
89. Lei, Y., Yalin, W., Paul, M. T., Vaibhav, A. N. & Jieping, Y. Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. Neuroimage 61, 622–632 (2012).
90. Yanbao, X., Siddharth, B., Shripirasad, R. D., Kevin, O. M. & Jimeng, S. RAIM: Recurrent attentive and intensive model of multimodal patient monitoring data. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2565–2573 (Association for Computing Machinery, 2018).
91. Yoxue, H., Mohd, U., Jun, Y., Hossain, M. S. & Ahmed, G. Recurrent convolutional neural network based multimodal disease risk prediction. Future Gener. Comput. Syst. 92, 76–83 (2019).
92. Shinichi, G. et al. Artificial intelligence-enabled fully automated detection of cardiac amyloidosis using electrocardiograms and echocardiograms. Nat. Commun. 12, 2726 (2021).
93. Rui, Y. et al. Richer fusion network for breast cancer classification based on multimodal data. BMC Med. Inform. Decis. Mak. 21, 134 (2021).
94. Li, Y., Wang, H. & Luo, Y. A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports. 2020 IEEE Int. Conf. Bioinform. Biomedicine (BIBM) 2020 1999–2004 (2020).
95. Sara Bersche, G. et al. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: A retrospective analysis of electronic medical records data. BMC Med. Inform. Decis. Mak. 18, 44 (2018).
96. Chowdhury, S., Zhang, C., Yu, P. S. & Luo, Y. Mixed pooling multi-view attention autoencoder for representation learning in healthcare. Preprint at https://doi.org/10.1101/2150199 (2020).
97. Ilan, S. et al. An unsupervised learning approach to identify novel signatures of health and disease from multimodal data. Genome Med. 12, 7 (2020).
98. Chowdhury, S., Zhang, C., Yu, P. S. & Luo, Y. Med2Meta: Learning representations of medical concepts with meta-embeddings. HEALTHINF 2020, 369–376 (2020).
99. Subramanian, V. Do MN, Syeda-Mahmood T. Multimodal fusion of imaging and genomics for lung cancer recurrence prediction. IEEE 17th International Symposium on Biomedical Imaging (ISBI), 804–808 (2020).
100. Michele, D. et al. Combining heterogeneous data sources for neuroimaging based diagnosis: Re-weighting and selecting what is important. Neuroimage 195, 215–231 (2019).
101. Yiwen, M., William, S., Michael, K. O. & Corey, W. A. Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression status in diffuse low- and high-grade gliomas. J. Neurooncol 142, 299–307 (2019).
102. Namyoung, P. et al. Predicting acute kidney injury in cancer patients using heterogeneous and irregular data. PLoS One 13, e0198839 (2018).
103. Wei Liang, T., Chee Kong, C., Sim Heng, O. & Alvin Choong Meng, N. Ensemble-based regression analysis of multimodal medical data for osteoarthritis diagnosis. Expert Syst. Appl. 80, 811–819 (2019).
104. Yiming, X. et al. Explainable dynamic multimodal variational autoencoder for the prediction of patients with suspected central precocious puberty. IEEE J. Biomed. Health Inform. 26, 1362–1373 (2021).
105. Alan, D. K. et al. Mixture model framework for traumatic brain injury prognosis using heterogeneous clinical and outcome data. IEEE J. Biomed. Health Inform. 25, 1321–1329 (2021).
106. Casper, R. et al. Preoperative risk stratification in endometrial cancer (ENDORISK) by a Bayesian network model: A development and validation study. PLoS Med. 17, e1003111 (2020).
107. Tommaso, G. et al. SARS-COV-2 comorbidity network and outcome in hospitalized patients in Crema, Italy. PLoS One 16, e0248498 (2021).
176. Xue, Y., Klabjan, D. & Luo, Y. Mixture-based multiple imputation model for clinical data with a temporal dimension. 2019 IEEE International Conference on Big Data (Big Data) 2019 245–252 (IEEE, 2019).
177. Cao, W. et al. Brits: Bidirectional Recurrent Imputation for Time Series. NeurIPS, 31 (2018).
178. Luo, Y. Evaluating the state of the art in missing data imputation for clinical data. Brief. Bioinform. 23, bkaa489 (2022).
179. Zhao, Q., Adeli, E. & Pohl, K. M. Training confounder-free deep learning models for medical applications. Nat. Commun. 11, 1–9 (2020).
180. Luo, Y. & Mao, C. ScanMap: Supervised confounding aware non-negative matrix factorization for polygenic risk modeling. In Machine Learning for Healthcare Conference; 2020. PMLR 27–45 (2020).
181. Kalavathy, R. & Suresh, R. M. Pharmacovigilance from electronic medical records to report adverse events. J. Chem. Pharm. Sci. 2015, 188–191 (2015).
182. Luo, Y. et al. Natural language processing for EHR-based pharmacovigilance: A structured review. Drug Saf. https://doi.org/10.1007/s40264-017-0558-6 (2017).
183. Segura Bedmar, I., Martínez, P. & Herrero Zazo, M. Semeval-2013 task 9: Extraction of drug–drug interactions from biomedical texts (ddiextraction 2013). 2013: Association for Computational Linguistics (2013).
184. Hammann, F. & Drewe, J. Data mining for potential adverse drug–drug interactions. Expert Opin. Drug Metab. Toxicol. 10, 665–671 (2014).
185. Donna, M. F. et al. American geriatrics society 2019 updated AGS beers criteria for potentially inappropriate medication use in older adults. J. Am. Geriatr. Soc. 67, 674–694 (2019).
186. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 366, 447–53. (2019).
187. Wang, H. et al. Using machine learning to integrate sociobehavioral factors in predicting cardiovascular-related mortality risk. Stud. Health Technol. Inf. 264, 433–437 (2019).
188. Christof, E. Open source software in industry. IEEE Softw. 25, 52–53 (2008).
189. Robert, M. S. Why develop open-source software? The role of non-pecuniary benefits, monetary rewards, and open-source licence type. Oxf. Rev. Economic Policy 23, 605–619 (2007).
190. Luo, Y., Wunderink, R. G. & Lloyd-Jones, D. Proactive vs reactive machine learning for potentially inappropriate medication use in older adults. JAMA 327, 623–624 (2022).
191. Ke, G. et al. LightGBM: A highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 30 (2017).
192. Derara Duba, R., Taye Girma, D., Achim, I. & Woruk Gachena, N. Diagnosis of diabetes mellitus using gradient boosting machine (Lightgbm). Diagnostics 11, 1714 (2021).
193. Xiaolei, S., Mingxi, L. & Zeqian, S. A novel cryptocurrency price trend forecasting model based on LightGBM. Financ. Res. Lett. 32, 101084 (2020).
194. Dongzi, J., Yiqin, L., Jiancheng, Q., Zhe, C. & Zhongshu, M. SwiftIDS: Real-time anomaly detection system based on LightGBM and parallel intrusion detection mechanism. Comput. Secur. 97, 101984 (2020).
195. Wilkinson, M. D. et al. The FAIR guiding principles for scientific data management and stewardship. Nature 3, 160018 (2016).
196. Weissler, E. H., et al. The role of machine learning in clinical research: Transforming the future of evidence generation. Trials. 22, 537 (2021).
197. Pratik, S. et al. Artificial intelligence and machine learning in clinical development: A translational perspective. npj Digital Med. 2, 69 (2019).
198. Inna, K. & Simeon, S. Interpretability of machine learning solutions in public healthcare: The CRISP-ML approach. Front. Big Data 4, 660206 (2021).
199. Marx, V. Method of the Year: Spatially resolved transcriptomics. Nat. Methods 18, 9–14 (2021).
200. Zeng, Z., Li, Y., Li, Y. & Luo, Y. Statistical and machine learning methods for spatially resolved transcriptomics data analysis. Genome Biol. 23, 1–23. (2022).
201. Longo, S. K., Guo, M. G., Ji, A. L. & Khavari, P. A. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. Nat. Rev. Genet. 22, 627–644 (2021).
202. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. Ann. Intern. Med. 169 467–473 (2018).

ACKNOWLEDGEMENTS
The study is supported in part by NIH Grants U01TR003528 and R01LM013337 to Y.L., and AG073323 to F.C.

AUTHOR CONTRIBUTIONS
A.K. led the review process, performed data extraction, performed the computation analysis, figure generation, writing, and dashboard creation. H.W., Y.L., S.D., and M.H. performed data extraction. Z.X. synthesized limitations of the studies. F.W. and F.C. performed proof reading and content curation. Y.L. conceived the review, oversaw the review process, and provided necessary feedback, proof reading, and content curation.

COMPETING INTERESTS
The authors declare no competing interests.

ADDITIONAL INFORMATION
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-022-00712-8.

Correspondence and requests for materials should be addressed to Yuan Luo.

Reprints and permission information is available at http://www.nature.com/reprints

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022