A template for the *arXiv* style

A Preprint

Randy *
Department of Biostatistics and Bioinformatics
University of Colorado
Aurora, CO, USA
xin.2.jin@cuanschutz.edu

August 22, 2022

Abstract

The traditional predictive modeling approach mainly relies on global inference and universal modelling with all available data. However, such approach may overlook the cultural diversity and genetic heterogeneity for patients. using fewer but more similar data could get higher predictive performance than using overall available data. Recently, curving matching prediction methods, based on predictive modeling with similar matching donars, has been successfully applied in the medical data analysis. This curve matching prediction aquires the information of "people-like-me", with the nearest neighbors of predictive mean. The predictive mean matching can avoid the dataset noise and model mispecification, with respect to given metrics. Through exhaustive comparisons for specific target and the most similar matching donor-cohort with predictive mean, an assessment specifc to the given patient can help in identifying his similar patients. In this paper,

The model is implicit (Little and Rubin 2002), which means that there is no need to define an explicit model for the distribution of the missing values. Because of this, predictive mean matching is less vulnerable

In contrast, the imputations created by predictive mean matching follow the data quite nicely, even though the predictive mean itself is clearly off-target.

There are different strategies for defining the set and number of candidate donors. Predictive mean matching performs very badly when d is small and there are lots of ties for the predictors among the individuals to be imputed.

The reason is that the tied individuals all get the same imputed value in each imputed dataset when $d = 1$ (Ian White, personal communication). Setting d to a high value (say $n/10$) alleviates the duplication problem, but may introduce bias since the likelihood of bad matches increases.

Schenker and Taylor (1996), evaluated $d = 3$, $d = 10$ and an adaptive scheme. The adaptive method was slightly better than using a fixed number of candidates, but the differences were small. compared various settings for d, and found that $d = 5$ and $d = 10$ generally provided the best results.

***K*eywords** predictive mean matching · personalized data-driven prediction · multiple imputation · Mahalanobis distance · people-like-me

1 Introduction

Disease populations, clinical practice, and healthcare systems are constantly evolving. This can result in clinical prediction models quickly becoming outdated and less accurate over time. A potential solution is

*Use footnote for providing further information about author

to develop 'dynamic' prediction models capable of retaining accuracy by evolving over time in response to observed changes. 1

Most commonly, through exhaustive comparisons between a given patient and a cohort of existing patients, an assessment specifc to the given patient can help in identifying his similar patients. The purpose of "people-like-me" methods is to find the donors Most commonly, through exhaustive comparisons between a given patient and a cohort of existing patients, an assessment specifc to the given patient can help in identifying his similar patients.

s who agreed the most with each patient. Te result suggested that using fewer but more similar data could get higher predictive performance than using overall available data. David et al. [20] proposed an algorithm for the anomaly detection and characterization on the basis of the Euclidean distance

The results demonstrated that personalized predictive models showed a higher performance. Many previous studies usually calculated the patient similarity using single similarity measures (e.g., Euclidean distance, cosine distance, and Mahalanobis distance), and most of them did not take the importance of patient features into consideration while calculating the similarity. In this study, we aimed to investigate in depth the patient similarity in the following two aspects. One is using diferent similarity metrics for diferent types of feature data. Te other is assigning diferent weights (importance) to patient features when integrating feature similarities into a patient similarity

many studies have found secondary use such as patient trajectory modeling, disease inference and clinical decision support system. It is recommended to denoise data before building a global predictive model, which will be time consuming and challenging to represent and model. In this context, individualized predictive modeling based on patient similarity emerged and was shown to be adjustable for individual patients. Employing patient similarity helps to identify a precision cohort for an index patient, which will then be used to train a personalized model. Accordingly, when building a predictive model for an index patient, training samples are determined as "patients like me," instead of using all available training samples in a conventional way. "Patients like me" are selected from the training sample set on the basis of similarity between the index patient and each training sample. Of note, based on patient similarity, patients with noisy data are less likely to be selected as similar patients of an index patient for the reason of the less similarity between them. Patient similarity is usually measured by considering information on demographics, disease history, comorbidities, laboratory tests, hospitalizations, treatment, and pharmacotherapy. Such data are easily extracted from the EMR for tens of millions of patients [13]. In this study, we defned a patient as a vector in a d-dimensional feature space. Ten, a multi-dimensional approach to estimate patient similarity was proposed. To demonstrate the efectiveness of the proposed similarity measure, the most similar patients were retrieved to build personalized models to predict the diabetes status of a given patient

Four methods has been proposed by Andridge and Little(2010) Hot deck imputation:

1. "The chosen threshold" Choose a threshold, and take all i for which as candidate donors for imputing j. Randomly sample one donor from the candidates, and take its yi as replacement value.
2. "The nearest neighbor" , i.e., the case i for which |yi yj | is minimal as the donor. This is known as "nearest neighbor hot deck," "deterministic hot deck" or "closest predictor."
3. "Sampling from k neighbors" Find the d candidates for which |yi yj | is minimal, and sample one of them. Usual values for d are 3, 5 and 10. There is also an adaptive method to specify the number of donors (Schenker and Taylor, 1996).
4. "Sampling from probability" Sample one donor with a probability that depends on |yi yj | (Siddique and Belin, 2008).

Existing approaches to modeling with predictive mean matching mainly rely on single predictive mean value on one single time point and a fixed number of candidate donors. These simplified modeling strategies may give a rise of several problems.

The GAMLSS method (Rigby and Stasinopoulos, 2005; Stasinopoulos et al., 2017) extends both the generalized linear model and the generalized additive model. A unique feature of GAMLSS is its ability to specify a (possibly nonlinear) model for each of the parameters of the distribution, thus giving rise to an extremely flexible toolbox that can be used to model almost any distribution. The gamlss package contains over 60 built-in distributions. Each distribution comes with a function to draw random variates, so once the gamlss model is fitted, it can also be used to draw imputations.

Various metrics are possible to define the distance between the cases. The predictive mean matching metric was proposed by Rubin (1986) and Little (1988). This metric is particularly useful for missing data applications because it is optimized for each target variable separately.

The predicted value only needs to be a convenient one-number summary of the important information that relates the covariates to the target. Calculation is straightforward, and it is easy to include nominal and ordinal variables.

The work presented here expands our previous preliminary study, as we 1) further assessed the adequacy of other existing embeddings for modeling medical concept dependence, 2) leveraged the similarity model by considering each diagnosis of ciliopathy as index (as opposed to using average similarity with all diagnosed patients) to take into account the high heterogeneity of ciliopathies, and 3) applied the developed model to two large-scale unbalanced datasets containing approximately 10,000 and 60,000 controls with kidney manifestations in the clinical data warehouse

8. Ng K, Sun J, Hu J, Wang F. Personalized predictive modeling and risk factor identifcation using patient similarity. AMIA Summits Transl Sci Proc. 2015;2015:132–6.
9. Whellan DJ, Ousdigian KT, Alkhatib SM, Pu W, Sarkar S, Porter CB, Pavri BB, O'Connor CM, Investigators PS. Combined heart failure device diagnostics identify patients at higher risk of subsequent heart failure hospitalizations: results from PARTNERS HF (program to access and review trending information and evaluate correlation to symptoms in patients with heart failure) study. J Am Coll Cardiol. 2010;55(17):1803–10.
10. Sepanski RJ, Godambe SA, Mangum CD, Bovat CS, Zaritsky AL, Shah SH. Designing a pediatric severe sepsis screening tool. Front Pediatr. 2014;2(56):56.
11. Wu J, Roy J, Stewart W. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med Care. 2010;48(6 Suppl):S106.

There are different strategies for defining the set and number of candidate donors. Setting d = 1 is generally considered to be too low, as it may reselect the same donor over and over again. Predictive mean matching performs very badly when d is small and there are lots of ties for the predictors among the individuals to be imputed. The reason is that the tied individuals all get the same imputed value in each imputed dataset when d = 1 (Ian White, personal communication). Setting d to a high value (say n/10) alleviates the duplication problem, but may introduce bias since the likelihood of bad matches increases.

Schenker and Taylor (1996) evaluated d = 3, d = 10 Morris et al. (2014) compared various Kleinke (2017) found that d = 5 may be too high for sample size lower than n = 100, and suggested setting d = 1 for better point estimates for small samples. Gaffert et al. (2016) explored scenarios in which candidate donors have different probabilities to be drawn, where the probability depends on the distance between the donor and recipient cases.

Instead a closeness parameter needs to be specified, and this was made adaptive to the data. An advantage of using all donors is that the variance of the imputations can be corrected by the Parzen correction, which alleviates concerns about insuffi- cient variability of the imputes. Their simulations showed that with a small sample (n = 10), the adaptive method is clearly superior to methods with a fixed donor pool.

an adaptive method for setting d could improve small sample behavior. Meanwhile, the number of donors can be changed through the donors argument.

Because of the bias, the coverage is lower than nominal. For missing x the bias is much smaller. Setting d to a lower value, as recommended by Kleinke (2017), improves point estimates, but the magnitude of the effect depends on whether the missing values occur in x or y. For the sample size n = 1000 predictive mean matching appears well calibrated for d = 5 for missing data in y, and has slight undercoverage for missing data in x. Note that Table 3.3 in the first edition of this book presented incorrect information because it had erroneously imputed the data by norm instead of pmm.

The traditional method does not work for a small number of predictors. Heitjan and Little (1991) report that for just two predictors the results were "disastrous." The cause of the problem appears to be related to their use

The method is robust against misspecification of the imputation model, yet performs as well as theoretically superior methods. In the context of missing covariate data, Marshall et al. (2010a) concluded that predictive mean matching "produced the least biased estimates and better model per- formance measures."

The method works best with large samples, and provides imputations that possess many characteristics of the complete data. Predictive mean matching cannot be used to extrapolate beyond the range of the data, or to interpolate within the range of the data if the data at the interior are sparse. Also, it may not perform well with small datasets. Bearing these points in mind, predictive mean matching is a great all-around method with exceptional properties.

there are two reasons to move beyond the predictive distance used in PMM and investigate an alternative metric. Firstly, PMM requires users of curve matching to select a particular future time point to base the matches on (e.g. 14 months of age). In some cases, it may be difficult to choose this time point, especially when the 'future' is more vaguely defined as a time interval.[4] Secondly, the predictive distance may make the matches look unconvincing. The trajectories of the selected donors may all be close to the prediction for the target child at 14 months, but this does not imply that the histories are identical. After all, different profiles may lead to the same predicted value.

Consequently, the curves of some of the matches may be quite far from the curve of the target child. Some users of curve matching feel that such discrepancies are undesirable, as these matches do not appear to be people-like-me.[4] It is useful to investigate these shortcomings not only for improving growth prediction but also for other applications of multiple imputation, such as patient recovery after an operation, prediction of longevity, and decision-making when more than one treatment is available. [3]

2 methods

Therefore, the information of these children at a later age is available. The first step is to fit a linear regression model on the donor database. Then, this model is used to predict the values for all donors and for the target at a certain point in the future, for example at 14 months. Finally, the distance between the predicted value of each of the donors and the predicted value of the target is calculated, which is referred to as the predictive distance. A number of donors – usually five - with the smallest predictive distance are selected as the best matches. Their growth curves are then plotted and point estimates can be calculated by averaging the measurements. The growth patterns of the matched children thus suggest how the target child might develop in the future.

the practical implementation and use of curve matching can in theory be improved by combining the predictive distance with another distance measure, thus creating a "blended distance" measure. Such a blended metric would take into account historical similarity between the donors and the target. For example, when blending the predictive distance with the Mahalanobis distance, more weight is given to similarities between units in the full predictor space. This would theoretically lead to the selection of donors with profiles more similar to the target, and therefore to the selection of true people-like-me. The objective of this study is to implement such a blended distance measure and to investigate its properties, blend ratio, and the validity of its resulting inferences.

So instead of using one predictive distance at particular future time point. we use multiple time; The PD is the distance between the predicted value of a donor and the predicted value of the target at a particular future time point. The MD is defined as the distance between two N dimensional points scaled by the variation in each component of the point.

3 Headings: first level

as the profiles of the matched donors can substantially differ from the profile of the target. similarity between the curves of the donors and the target can be taken into account by combining the predictive distance with the Mahalanobis distance.

. The results show that blending towards the Mahalanobis distance leads to worse performance in terms of bias, coverage, and predictive power. Simulation study II evaluates the blended metric in a setting where a single value is imputed. The results show that a property of blending is the bias-variance trade off. Giving more weight to the Mahalanobis distance leads to less variance in the imputations, but less accuracy as well.

We used the following steps to train and test both PLM and LMM prediction approaches: (1) build the approach using the training data, (2) examine prediction performance and tune the approach using the training data (i.e., within-sample testing), and (3) test the accuracy and precision of each approach using the testing data (i.e., out-of-sample testing). We compared performance of PLM and LMM predictions in terms of accuracy and precision across all individuals and all timepoints.

When no more than 30% of the whole training sample (i.e., 3000 samples) were used to build the models, all three personalized predictive models outperformed the corresponding traditional models, which were built on randomly selected training samples of the same size as the personalized models

We applied a form of predictive mean matching to determine the relative weights for each matching variable using the following steps.28 First, we imputed a 365- day TUG value for each patient in the training dataset via the brokenstick package in R (because data were collected at irregular timepoints).29 Next, we created a linear model to estimate the imputed 365-day TUG value using our matching variables of interest We then used this linear model to estimate the 365-day TUG value for (a) each patient in the training dataset and (b) the index patient. Finally, the patients from the training dataset with the closest predicted 365-day TUG value to the index patient were selected as matches; we used 35 patient matches based on our previous work.12

We modeled the observed TUG data from the matching patient records to form the index patient's predicted TUG recovery trajectory. We used Generalized Additive Models for Location, Scale, and Shape (GAMLSS) to flexibly model the median, variance, and skewness of TUG recovery from postoperative days 1-425.30 The GAMLSS model included a cubic spline smoother with 3 degrees of freedom for the location parameter and 1 degree of freedom for the shape parameter.

A previous study suggested that in personalized medicine, using patient similarity in data-driven analysis of patient cohorts will signifcantly assist physicians to make informed decisions and choose the most appropriate clinical trial

(why it is reasonable and useful)

- **what type of study**: The EPIC Observational Study is a prospective, multicenter, observational longitudinal study
- **multiple center**: The data was collected through the Cystic Fibrosis Foundation Patient Registry (CFFPR)
- **already in use**: The design of the EPIC study has been previously reported [28,29]. Including 59 centers
- **exclusive informations**: Include demographic and risk factor data, respiratory cultures, clinical encounter information, and clinical outcomes, such as bacterial infections or pulmonary exacerbations
- **what is the goal**:
- **our goal**:

3.0.1 Subjects

- **check for time!!**: 1772 participants enrolled between *2004 and 2006*. The data cut-off for the current analysis was *December 2013*.
- The enrolled children with a confirmed diagnosis of CF before the age of 12 without exposure of *Pseudomonas aeruginosa*
- Based on either never infected or culture negative for at least 2 years); with respiratory cultures negative for *Pseudomonas aeruginosa*
- Either no prior isolation of Pseudomonas aeruginosa, or at least a two-year period with no isolation of P. aeruginosa (Treggiari et al., 2009).
- The beginning of follow-up was defined as an individual's earliest registry entry or pulmonary exacerbation
- a registry entry could have occurred before their enrollment in the EPIC Observational Study
- End of follow-up in the study was defined as a patient's latest CF registry encounter or pulmonary exacerbation.
- **sample size**: Of the **1772** children enrolled, **76497** visit observations in the EPIC OBS study
- we identified **1325** individuals with usable data
- 1325 for the final data analysis
- **how many** missing or lost to-follow-up with time-varying covariates and outcomes

- % had one survey record, % had a second, and % had a third, and the remaining % had **#th** surveys.
- Total of **# obs** in **# subject** individuals were identified after eliminating registration age over 5
- Total of **# obs** in **# subject** individuals were identified after eliminating follow up time less than 5 years
- Total of **# obs** in **# subject** individuals were identified after eliminating avaiable visit times less than 10
- **testing/training**: randomly

3.0.2 Table1 Information

- half of the cohort was female (49%)
- most were Caucasian (95%)
- Our genotype of interest was defined by the number of confirmed $\Delta F508$ (`F508del`) mutations (`0`: without mutation, `1`: with one allele mutated, or `2`: with both alells mutated), where `0` included both those with both alleles missing or those with unknown status with respect to this mutation.
- Homozygosity of $\Delta F508$ ($\Delta F508/\Delta F508$)
- the most common CF-causing mutation (Cystic Fibrosis Foundation,

2016) is associated with increased risk of infection and lower lung function in CF (Rosenfeld et al., 2012, Sanders et al., 2010).

- the majority were F508del homozygous (%),
- the median weight percentile at enrollment was 36.7 **(IQR:)**.
- The median age at enrollment was 5.8 **(IQR:)** years
- median length of follow-up 7.8 **(IQR:)** years

3.0.3 Tools

the **brokenstick v 2.0.0** package. The package contains tools to fit the broken stick model to data, export the fitted model's parameters, create imputed values of the model, and predict broken stick estimates for new data.

3.1 brokenstick model

Substantive researchers often favour repeated measures over the use of linear mixed models because of their simplicity.

For example, we can easily fit a subject-level model to predict future outcomes conditional on earlier data with repeated measures data.

While such simple regression models may be less efficient than modelling the complete data (Diggle, Heagerty, Liang, and Zeger 2002, Sec. 6.1),

The broken stick model requires a specification of a sensible set of time points at which the measurements ideally should have been taken.

- For each subject, the model predicts or imputes hypothetical observations at those times
- so the substantive analysis applies to the repeated measures instead of the irregular data.

applications for brokenstick model:

- to approximate individual trajectories by a series of connected straight lines; • to align irregularly observed curves to a joint age grid;
- to impute realisations of individual trajectories;
- to estimate the time-to-time correlation matrix;
- to predict future observations.

provide the examples:

- the points are close to the scheduled ages (indicated by vertical lines), especially in the first half-year.
- bservation times vary more for older children.
- several children have one or more missing visits.
- some children had fairly close visits and vary periodically
- special cases of dropped out after certain time

3.2 people-like-me methods

Assumptions:

- The trajectory between break ages follows a strict linear functino. This assumption may fail for processes that are convex or concave in time. For example, human height growth in centimeters growth is concave, so setting breakpoints far apart results introduces systematic model bias. Modeling height SDS instead of raw height will prevent this bias.
- The broken stick estimates follow a joint multivariate normal distribution. As this assumption may fail for skewed measurements, it could be beneficial to transform the outcomes so that their distribution will be closer to normal.
- The data are Missing at Random (MAR) given the outcomes from all subjects at all observation times. This assumption is restrictive in the sense that missingness may only depend on the observed outcomes, and not on covariates other than time.

At the same time, the assumption is liberal in the sense that the missingness may depend on future outcomes. While this MAR-future assumption is unusual in the literature on drop-out and observation time models, it is a sensible strategy for creating imputations that preserve relations over time, especially for intermittent missing data. Of course, the subsequent substantive analysis on the imputed data needs to be aware of the causal direction of time.

3.2.1 KR fucntion???

The brokenstick package provides another alternative, the Kasim- Raudenbush (KR) sampler (Kasim and Raudenbush 1998). The method simulates draws from the posterior distributions of parameters from a two-level normal model with heterogeneous within-subject variances. The speed of the Kasim-Raudenbush sampler is almost insensitive to the number of random effects and depends primarily on the total number of iterations and somewhat on sample size.

4 Discussion

5 Reference

6 Supplementary

7 Examples of citations, figures, tables, references

You can insert references. Here is some text (Kour and Saabne 2014b, 2014a) and see Hadash et al. (2018).

The documentation for `natbib` may be found at

You can use custom blocks with LaTeX support from **rmarkdown** to create environment.

`http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf%7D`

Of note is the command `\citet`, which produces citations appropriate for use in inline text.

You can insert LaTeX environment directly too.

```
\citet{hasselmo} investigated\dots
```

produces

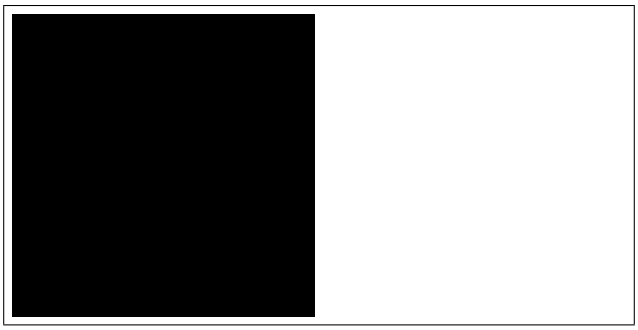

Figure 1: Sample figure caption.

Table 1: Sample table title

Part		
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

Hasselmo, et al. (1995) investigated...

`https://www.ctan.org/pkg/booktabs`

7.1 Figures

You can insert figure using LaTeX directly.

See Figure 1. Here is how you add footnotes. [^Sample of the first footnote.]

But you can also do that using R.

```
plot(mtcars$mpg)
```

You can use **bookdown** to allow references for Tables and Figures.

7.2 Tables

Below we can see how to use tables.

See awesome Table~1 which is written directly in LaTeX in source Rmd file.

You can also use R code for that.

```
knitr::kable(head(mtcars), caption = "Head of mtcars table")
```

Table 2: Head of mtcars table

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

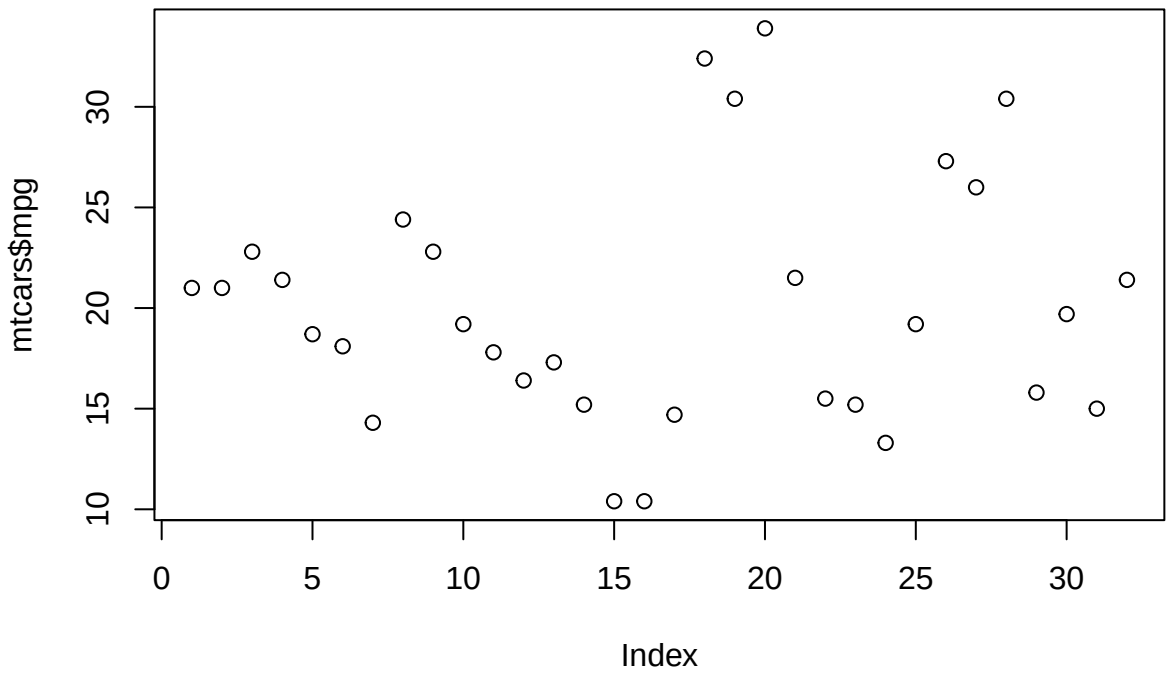

Figure 2: Another sample figure

7.3 Lists

- Item 1
- Item 2
- Item 3

Hadash, Guy, Einat Kermany, Boaz Carmeli, Ofer Lavi, George Kour, and Alon Jacovi. 2018. "Estimate and Replace: A Novel Approach to Integrating Deep Neural Networks with Existing Applications." *arXiv Preprint arXiv:1804.09028*.

Kour, George, and Raid Saabne. 2014a. "Fast Classification of Handwritten on-Line Arabic Characters." In *Soft Computing and Pattern Recognition (SoCPaR), 2014 6th International Conference of*, 312–18. IEEE.

———. 2014b. "Real-Time Segmentation of on-Line Handwritten Arabic Script." In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, 417–22. IEEE.