Explaining Multiple Instances Counterfactually:User Tests of Group-Counterfactuals for XAI

Greta Warren[1,2,3], Eoin Delaney[4], Christophe Guéret[5],
and Mark T. Keane[1,2(✉)]

[1] School of Computer Science, University College Dublin, Dublin, Ireland
grwa@di.ku.dk

[2] Insight Centre for Data Analytics, University College Dublin, Dublin, Ireland
mark.keane@ucd.ie

[3] Department of Computer Science, University of Copenhagen,
Copenhagen, Denmark

[4] Oxford Internet Institute, Oxford University, Oxford, UK
eoin.delaney@oii.ox.ac.uk

[5] Accenture Labs, Dublin, Ireland
christophe.gueret@accenture.com

Abstract. Counterfactual explanations have become a major focus for post-hoc explainability research in recent years, as they seem to provide good algorithmic recourse solutions, people can readily understand them, and they may meet legal regulations (such as GDPR in the EU). However, this large literature has only addressed the use of counterfactual explanations to explain single predictive-instances. Here, we explore a novel use case in which groups of similar instances are explained in a collective fashion using "group counterfactuals" (e.g., to highlight a repeating pattern of illness in a group of patients). Group counterfactuals potentially provide broad explanations covering multiple events/instances. A novel case-based, group-counterfactual algorithm is proposed to generate such explanations and a user study is also reported to test the psychological validity of the algorithm.

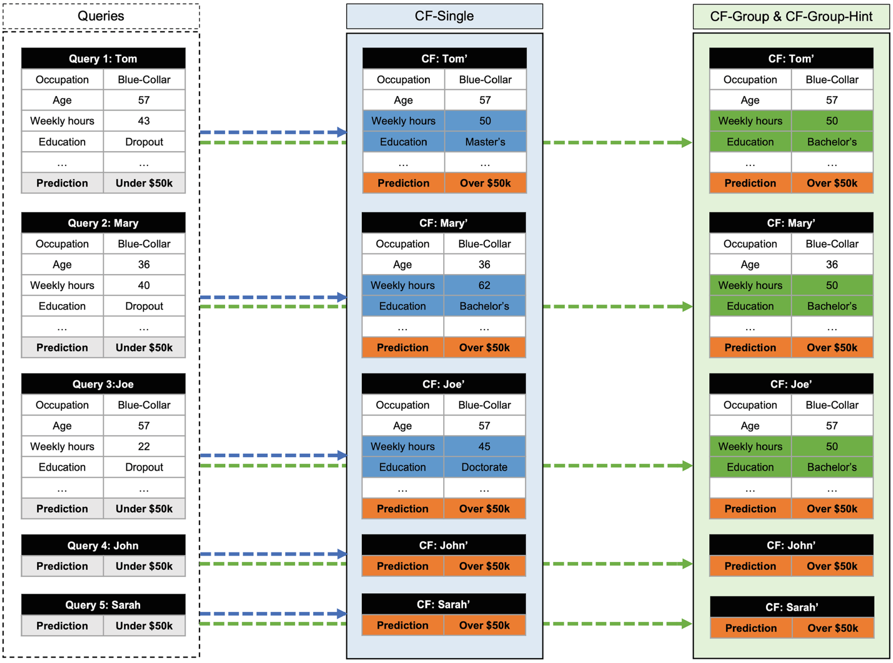
Keywords: XAI · Explainability · Counterfactuals · User-Centered

1 Introduction

In recent years, the literature on eXplainable AI (XAI) has focused significantly on the use of counterfactual explanations to explain the predictions of opaque machine learning (ML) models [14,19,22,24,30], as they show what changes to input-features can alter a model's output decisions (e.g., "if only the bank customer had applied for a lower loan amount of $10,000, their loan application would have been approved"). Interest in counterfactuals as explanations has been boosted by arguments made in philosophy and psychology about the formal

J. A. Recio-Garcia et al. (Eds.): ICCBR 2024, LNAI 14775, pp. 206–222, 2024.
https://doi.org/10.1007/978-3-031-63646-2_14

analysis of causality [28] and people's causal thinking [4,30], respectively. Furthermore, legal analyses have suggested that counterfactual explanations comply with General Data Protection Regulation (GDPR) requirements [39], leading to their extensive use in algorithmic recourse [19].



Queries

Query 1: Tom	
Occupation	Blue-Collar
Age	57
Weekly hours	43
Education	Dropout
...	...
Prediction	Under $50k

Query 2: Mary	
Occupation	Blue-Collar
Age	36
Weekly hours	40
Education	Dropout
...	...
Prediction	Under $50k

Query 3:Joe	
Occupation	Blue-Collar
Age	57
Weekly hours	22
Education	Dropout
...	...
Prediction	Under $50k

Query 4: John	
Prediction	Under $50k

Query 5: Sarah	
Prediction	Under $50k

CF-Single

CF: Tom'	
Occupation	Blue-Collar
Age	57
Weekly hours	50
Education	Master's
...	...
Prediction	Over $50k

CF: Mary'	
Occupation	Blue-Collar
Age	36
Weekly hours	62
Education	Bachelor's
...	...
Prediction	Over $50k

CF: Joe'	
Occupation	Blue-Collar
Age	57
Weekly hours	45
Education	Doctorate
...	...
Prediction	Over $50k

CF: John'	
Prediction	Over $50k

CF: Sarah'	
Prediction	Over $50k

CF-Group & CF-Group-Hint

CF: Tom'	
Occupation	Blue-Collar
Age	57
Weekly hours	50
Education	Bachelor's
...	...
Prediction	Over $50k

CF: Mary'	
Occupation	Blue-Collar
Age	36
Weekly hours	50
Education	Bachelor's
...	...
Prediction	Over $50k

CF: Joe'	
Occupation	Blue-Collar
Age	57
Weekly hours	50
Education	Bachelor's
...	...
Prediction	Over $50k

CF: John'	
Prediction	Over $50k

CF: Sarah'	
Prediction	Over $50k

Fig. 1. Sample Queries with Single or Group Counterfactual Explanations. Five summary queries are shown paired with single or group counterfactual explanations (n.b., John and Sarah details are finessed). Note that target-values for Weekly Hours and Education in the single counterfactuals vary in each explanation, whereas those for the group counterfactuals are the same in each explanation; critically, all of the feature-differences in these pairings create valid counterfactuals that flip the outcome class.

However, counterfactual XAI has been criticised for not paying sufficient attention to suitable use cases, to determining when counterfactual explanations work well [3,21]. Indeed, mixed results from recent user studies may arise from inappropriate usages (see e.g., [7,8,26,38,40]). Most counterfactual use-cases assume scenarios in which a *single* prediction for a *single instance* is explained using a *single* explanation (e.g., a loan application). However, AI systems also make *multiple* predictions for *similar instances with the same outcome*, predictive-instances that could be explained in a grouped way using the "same" counterfactual explanation (i.e., feature-differences with identical target-values, see Fig. 1). The current work came from end-user feedback trialling XAI models for disease prediction [32,36]. When we showed farmers disease predictions for a cow explained by a counterfactual (e.g., "if this animal had a lower cell count, it would be healthy"), they said that typically several animals fall ill at a time and explaining the whole group together would be better (e.g., "if these four animals had lower cell counts and were younger then they would be healthy"); that is, they felt a group counterfactual could surface patterns of disease in the herd, leading to better disease interventions (e.g., isolating animals).

1.1 When Explanations for Groups Might Help

Different stakeholders may have different requirements and desiderata for explanations [27]. Previous work on counterfactual explanations has focused on end-users affected by automated decisions (e.g., a bank customer who has applied for a loan), where a single explanation or a set of diverse explanations are provided for a single prediction [31,39]. However, other stakeholder users may want explanation for patterns of predictions for similar queries. For example, the farmer wanting one explanation for a group of sick animals (rather than a different counterfactual for each animal) or a bank manager might want single explanation for a group of customers to determine if they are being discriminated against by the AI system. For example, in our use-case we assume the stakeholder is a risk analyst assessing groups of bank customers to determine if they have been correctly treated in automated decisions, a real use-case in the home loan sector. Hence, we used the Adult (Census) dataset [11] to build a classifier that predicted whether individuals earned under or over $50,000 in annual income (see Fig. 1). We note that people seem to have a reasonable level of knowledge about this domain; without training, they are quite accurate in their classifications (e.g., ∼70–80% accurate; see later user tests).

Our novel case-based algorithm for group counterfactuals (**Group-CF**) explains binary classifications for similar individuals from the same class using a group counterfactual with common feature-values (see Fig. 1). In contrast, a traditional single-counterfactual method (**Single-CF**) explains each instance using a different counterfactual; that is, "if Tom worked *50* hours per week instead of 43, he would have earned over $50k", "if Mary worked *62* h per week instead of 40, she would have earned over $50k", "if Joe worked *45* h per week instead of 22, he would have earned over $50k" and so on. Group-CF generates a group counterfactual re-using the same target-value in the feature-differences for instances; that is, "if Tom worked *50* h per week instead of 43, he would have earned over $50k", "if Mary worked *50* h per week instead of 40, she would have earned over $50k", "if Joe worked *50* hours per week instead of 22, he would have earned over $50k" and so on. Note, here, we do not consider how pools of queries are selected to be explained; they could identified by the user or by clustering similar cases to present to the user. For present purposes, we randomly selected a seed-query, then created a pool of k-nearest instances (usually, $k = 5$), computing a group counterfactual for them.

1.2 Related Work

Single-counterfactual methods provide *local* (or quasi-local) explanations rather than *global* ones (that explain the whole dataset), in that they attempt to explain a circumscribed region of the decision space (sometimes called *cohort* explanations). Local explanations using single-counterfactuals for XAI have been studied extensively (see e.g., [18,21,37] for reviews), with probably 150+ distinct algorithms. However, work on group counterfactuals is recent and rare. Kanamori et al. [17] proposed Counterfactual Explanation Trees (CETs) that covered multiple instance predictions from decision trees as a"summary of local explanations". In a 2023 workshop paper, Warren, et al. [41] proposed the instance-based, CBR method for group-counterfactuals reported here. More recently, several other papers have formally described group counterfactuals along with optimisation methods for their computation [1,2,5,6]). Artelt & Gregoriades [1,2] compute group counterfactuals for a human resources dataset dealing with employee attrition; their group counterfactuals explain what needs to happen to prevent employees leaving an organisation (e.g., "If these employees had a salary increase of 20%, AND increased job satisfaction of 50%, attrition would be unlikely"). They also competitively tested their method against other methods (i.e., [17,41]), showing that it was more *cost effective* (in its sparsity) and *correct* (in instance coverage) [2]. In a related vein, hyperbox methods for forming general descriptions of collections of instances, called Actionable Recourse Summaries [34], Global Counterfactual Explanations [33], and Interpretable Regional Descriptors [9]) – have been advanced though they tackle somewhat different problems.

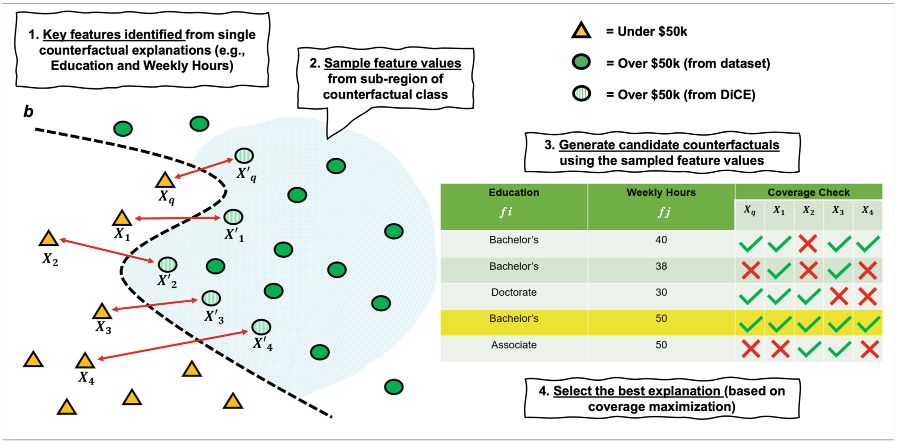
The rapid growth in counterfactual methods for explaining multiple instances underscores their potential utility in XAI. However, a major gap in this work is a failure to consider psychological validity. There are no existing user tests of group counterfactuals or methodologies proposed for testing them. Although these papers tend to agree on the form of group counterfactuals (i.e., stated as specific-value differences for a group of instances), we do not know if people can comprehend this type of explanation and if so, whether they are better than using several single-counterfactuals. However, there are hints in the cognitive literature that group-counterfactuals might work; good explanations tend to be more general [23,29] and people prefer broad scope explanations covering several observations [12,16,35]. For instance, explanations for disease diagnoses that account for three observed symptoms are judged to be better than those that account for just one [29]. A major contribution of this paper is a methodology for user testing the group-counterfactuals idea and its application in a concrete testing situation.

1.3 Paper Outline and Contributions

The present paper proposes a novel case-based algorithm for group counterfactuals (called Group-CF) and a first user test of group counterfactuals. As such, it makes two key contributions:

- *Algorithmic Development*: the development of a novel case-based counterfactual XAI algorithm that groups explanations of similar predictive-instances, along with a methodology for presenting these to end-users (see Sect. 2)
- *User Study*: the first user evaluation of the group-counterfactual idea, carefully designed to assess its impact on objective (i.e., accuracy) and subjective (i.e., confidence, satisfaction and trust) psychological measures of human understanding, relative to traditional single counterfactuals (see Sect. 3)

In the final section of the paper, we consider some of the caveats, limitations and future directions in this area that need to be addressed (Sect. 4).


1. Key features identified from single counterfactual explanations (e.g., Education and Weekly Hours)
2. Sample feature values from sub-region of counterfactual class
3. Generate candidate counterfactuals using the sampled feature values
4. Select the best explanation (based on coverage maximization)
= Under $50k
= Over $50k (from dataset)
= Over $50k (from DiCE)
b
X_q
X'_q
X_1
X'_1
X_2
X'_2
X_3
X'_3
X_4
X'_4

Education fi	Weekly Hours fj	Coverage Check				
		X_q	X_1	X_2	X_3	X_4
Bachelor's	40	✓	✓	✗	✓	✓
Bachelor's	38	✗	✓	✗	✓	✗
Doctorate	30	✓	✓	✓	✗	✗
Bachelor's	50	✓	✓	✓	✓	✓
Associate	50	✗	✗	✓	✓	✗

Fig. 2. A group of queries, $(X_q, X_1, X_2, ..., X_n)$, are classed as being on one side of a decision boundary (e.g., Under $50k). Single counterfactuals are generated for each of these points, forming several individual explanations revealing two feature differences commonly used (i.e., Education and Weekly Hours are $[f_i, f_j]$). New values for the identified features are sampled from a region of training instances in the contrasting class (i.e., Over $50k) and substituted into the query instances to create candidate group-counterfactuals (X_{sub_1}, X_{sub_2},..., X_{sub_n}) checked for validity, the best being chosen on coverage.

2 A Group-Counterfactual Algorithm: Group-CF

Figure 2 illustrates the main steps in the **Group-CF** algorithm developed to compute a group counterfactual for some selected pool of instances (see Algorithm 1 for formalization). The algorithm starts from a set of related training instances that have been correctly classified as being from the same class[1]; for example, a pool containing a query instance, X_q, and its nearest-like-neighbours

[1] It is not enough just to select training instances with class label c as the model may not agree with this label due to regularization of the classifier to prevent over-fitting.

$(X_1, X_2, ..., X_{n-1})$ (n.b., pool size is a hyperparameter, here set to 5). So, we adopt a simple solution to the pool selection issue mentioned earlier.

Taking these inputs, the four main steps of the method are: (i) *identifying key difference-features* by generating individual counterfactuals for each related instance and analysing the feature differences on which they rely, and then (ii) *sampling feature-values* for key difference-features from data-points in the contrasting class for (iii) *generating candidate group-counterfactuals* based on substituting these feature-values into the original instances before (iv) *selecting the best group-counterfactual* based on its valid coverage of all the instances in the pool. Note, this method does *not* generalise the individual counterfactuals for the instances, rather it uses them to guide feature-selection towards good group-counterfactuals (see https://github.com/e-delaney/group_cfe). As the method leverages class labels, it is supervised and not a form of unsupervised learning (e.g., clustering).

2.1 Step 1: Key-Feature Identification

Given a pool of to-be-explained instances (of size n) – a query, X_q, and its nearest-like-neighbours $(X_1, X_2, ...X_{n-1})$ – and an opaque black-box model, b, ensure that $b(X) = c$, where c is the predicted class. Then, generate individual counterfactuals for all instances in the pool, forming a set of counterfactual instances $(X'_q, X'_1, X'_2, ..., X'_{n-1})$ such that $b(X') = c'$. Any "traditional" counterfactual method can be used to generate these diverse counterfactuals; we used DiCE [31] due to its popularity, above baseline performance, and open source-code availability (n.b., DiCE, as a generative method, is known to periodically produce implausible counterfactuals; so, these were filtered out when selecting items in the user study). The feature labels that are altered in the individual counterfactual generation are counted across all counterfactuals to determine the most commonly used features. For example, Education and Weekly Hours emerge here as the most frequently used difference-features that flip the classification (see Figs. 1 and 2), so these would be chosen as the two key features (i.e., $[f_i, f_j]$) to modify in generating candidate group-counterfactuals (n.b., majority voting is just one way to do this). This step also identifies the direction-of-difference in the key features (e.g., increase/decrease for a continuous feature) that flips the classification, that is used in step 2.

2.2 Step 2: Sample Feature Values

Having identified the key features in the single counterfactuals, we need to perturb their values to create candidate group counterfactuals, from which we will select one that has a common set of feature-difference values for them all. To constrain possible selections, we sample from training data in a sub-region of the **contrasting class** (e.g., the Over $50k class). For example, for the Education feature, this sampling identifies candidate values such as Bachelor's, Doctorate or Associate's degree (see Fig. 2). Importantly, these sampled instances are known, valid data-points, and therefore, are more likely to yield feature-value transformations that result in valid counterfactuals. Obviously, it makes sense to reduce the size of this sub-region (e.g., only consider data points with same direction-of-difference such as a higher educational qualification than the to-be-explained instances or data-points that are within a certain distance). In addition, using feature-values from prototype-like instances in this sub-region works well (e.g., medoids from k-medoids clustering).

2.3 Step 3: Generating Candidates

The feature-values from the sub-region's data-points are interpolated into the original counterfactuals to create candidate group counterfactuals (i.e., X_{sub}; see Fig. 2). Next, each of the candidate counterfactuals is passed to the model to verify that they are indeed valid and predicted to be in the counterfactual class (coverage check).

2.4 Step 4: Selecting the Best Explanation

The feature-value substitutions that create the candidate group-counterfactuals are checked for validity and coverage. The validity check determines whether the feature-changes do indeed flip the classification of a given instance to the

Algorithm 1. Group-CF Method

```
Require: $b(.)$; to-be-explained black-box model
Require: $D_c$; Instances in the training data with class label $c$
Require: $D_{c'}$; Instances in the training data with class label $c'$
Require: $X_q$; Query instance with features $[f_1, f_2, ..., f_k]$ s.t. $b(X_q) = c$
Require: $n$; the size of the group (including the query, $X_q$)
  Retrieve $X_{NLN} \in D_c$, the Nearest Like Neighbour subset pool, $\{X_1, X_2, ..., X_{n-1}\}$, for the query are selected such that $b(X) = c \,\forall\, X \in X_{NLN}$.
  for $X \in \{X_q, X_1, X_1, ..., X_{n-1}\}$ do
      Generate individual CFE, $X'$, s.t. $b(X') = c'$.
      Note the feature change and the direction of change if applicable.
  end for
  The most commonly perturbed feature set, $[f_i, f_j]$, from the individual counterfactuals informs the features to-be-changed in the Group-CF.
  Randomly sample feature pairs $[f_i, f_j]$ from sub-region, $R$, of the training data in the counterfactual class $R \in D_{c'} \rightarrow samples$
  for $[f_i, f_j]$ in samples do
      for $X \in \{X_q, X_1, X_1, ..., X_{n-1}\}$ do
          substitute feature values with $[f_i, f_j] \rightarrow X_{sub}$
          If $b(X_{sub}) \neq c'$ Stop
      end for
      Return sample feature pair that maximises coverage,
  end for
  Stop
```

contrasting class (i.e., $b(X_{sub}) = c'$). The coverage check determines whether this classification change holds over all the original instances in the pool (i.e., the number of valid counterfactuals created when substituting the feature values $[f_i, f_j]$ into the original group of instances). The group counterfactual with the highest coverage is chosen to explain multiple instances (see Fig. 2). In this step, if this "best" candidate fails to cover all instances in the pool, then those that are not covered would be excluded from the original to-be-explained set (perhaps, to be explained as part of a different pool).

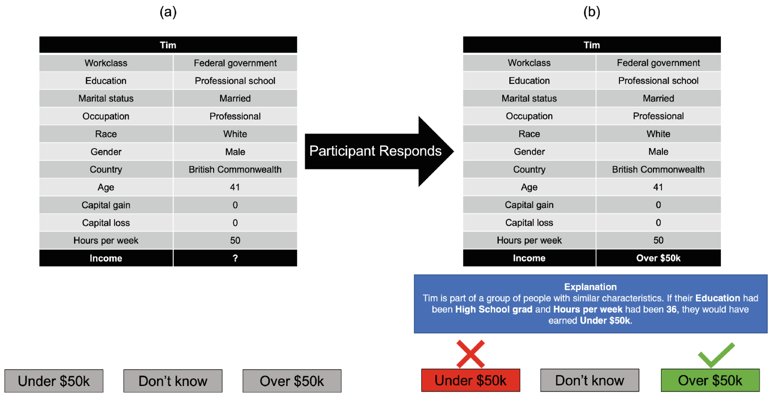
3 Group Counterfactuals: A User Study

Most current methods produce group-counterfactuals that look like those generated by our instance-based, Group-CF method. However, we do not know whether people can comprehend this type of counterfactual and whether they work better than several, single counterfactuals. Hence, we designed a user study to measure whether group counterfactuals improved people's understanding of an AI decision-making system compared to single counterfactuals. The study had two phases: a (i) *training phase*, in which people were shown instances from a dataset and asked to predict their outcomes before being shown the AI system's prediction along with an explanation (see Fig. 3), and a (ii) *testing phase*, in which people were shown instances and asked to predict their outcomes with no feedback or explanation as to their correctness (akin to solely getting part-a of Fig. 3 on its own). The main measure was *accuracy*, the proportion of instance-items correctly predicted in the testing phase (i.e., corresponding to the model's prediction). Subjective measures of confidence, satisfaction and trust were also recorded. Participants were shown 40 instances in each phase of the study (i.e., 80 unique items in total), with no overlap between the items in each phase. In the training phase, the 40 items were made up of eight 5-item groups of similar instances for which group counterfactuals were generated (in the relevant conditions). To control for possible order effects, the material sets were randomly re-ordered in each phase for each participant. Participants were assigned to one of three conditions – CF-Single, CF-Group, or CF-Group-Hint – that were matched in every respect except for the type of counterfactual explanations provided during the training phase (see Sect. 3.1 for details). Participants in the CF-Single condition were presented with classifications that were explained using diverse, single counterfactual explanations. Participants in the CF-Group and CF-Group-Hint conditions were presented with the same classifications, which were explained using group counterfactuals for related classified instances. The participants in the CF-Group-Hint condition received an additional "hint" along with each explanation that informed users that the individual belonged to a related group of people, to explicitly signal the commonality between the counterfactual instances.

So, this design tests whether the experience of seeing single/group counterfactuals in the training phase, improves their knowledge of the model's predictions in the test phase. This was a hard test of the group-counterfactual proposal, as we did not present the pool of queries together as a group in the test phase (i.e., queries with a common group counterfactual were randomly distributed in the 40 presented items). So, participants had to spontaneously notice that some items had common explanations. We chose this procedure as it seemed to, perhaps, better correspond to some real-life use-cases (e.g., where an analyst is sequentially assessing several customer-decisions and the model has determined in the background that some are part of similarly-treated pool). If we had presented the pool of queries together, the study could be criticised for making the group counterfactuals too obvious.

3.1 Method: Design and Procedure

The study had a 3 (Explanation: CF-Single vs CF-Group vs CF-Group-Hint) x 2 (Phase: Training vs Testing) mixed design with Explanation as a between-participant and Phase as a within-participant variable. The three Explanation conditions varied the counterfactual explanations provided in the training phase: *CF-Single* gave people instance-predictions explained using a single counterfactual unique to that instance (e.g., "If Joe's *Weekly hours* had been *45* and his *Education level* had been *Doctorate degree*, he would have earned Over $50k"; see Fig. 1), *CF-Group* gave people instance-predictions explained using group-counterfactuals with common target-values, implicitly grouping the similar instances in a given 5-item set (e.g., "If Joe's *Weekly hours* had been *50* and his *Education level* had been *Bachelor's degree*, he would have earned Over $50k"; see Fig. 1), and *CF-Group-Hint* presented people with the same group-counterfactuals as in CF-Group along with an explicit "hint" saying the instance was "part of a group of people with similar characteristics" (see Fig. 3). CF-Single



(a)

Tim	
Workclass	Federal government
Education	Professional school
Marital status	Married
Occupation	Professional
Race	White
Gender	Male
Country	British Commonwealth
Age	41
Capital gain	0
Capital loss	0
Hours per week	50
Income	**?**

(b)

Tim	
Workclass	Federal government
Education	Professional school
Marital status	Married
Occupation	Professional
Race	White
Gender	Male
Country	British Commonwealth
Age	41
Capital gain	0
Capital loss	0
Hours per week	50
Income	**Over $50k**

Explanation
Tim is part of a group of people with similar characteristics. If their **Education** had been **High School grad** and **Hours per week** had been **36**, they would have earned **Under $50k**.

Fig. 3. Sample Material from Training Phase of the User Study. The participant first sees (a) with the task of providing their own prediction for the item (3 options). After they respond, they are presented with (b), showing feedback on the correctness of their response along with an explanation (in this case for the CF-Group-Hint condition).

is essentially the control condition, with CF-Group and CF-Group-Hint being the experimental conditions, all three being matched in other respects.

At the beginning of the study, participants were informed that they would be testing an AI system designed to predict people's annual income from available information about them. Participants read detailed instructions about the task and provided informed consent. After completing practice trials for each phase of the study, participants progressed through both phases of the main task, and subsequently completed the subjective measures (i.e., Explanation Satisfaction and Trust scales). During the training phase, on each screen participants were presented with an instance without the class prediction and asked to make an *income* prediction from three options – Under $50k / Don't know / Over $50k – as shown in Fig. 3(a). Button order was randomised for each item, to prevent users from repeatedly selecting the same response. After making their prediction, users were shown feedback, with the AI model's prediction (correct answer) shown in green with a tick-mark, and the incorrect answer shown in red with a cross-mark. Figure 3(b) shows an example from the Group-CF-Hint condition where a correct prediction was made and explained using a group-counterfactual with a hint. After completing the 40 items in the training phase, people progressed to the testing phase, in which they were shown 40 further instances without outcomes shown and asked to choose one of the three options to make their prediction (as in part (a) of Fig. 3 with no further feedback). Here, after each response, participants rated their confidence in their prediction (using a 5-point Likert scale, from 1 (*Not at all confident*) to 5 (*Extremely confident*). In the testing phase, participants progressed through the 40 items, providing their predictions and confidence judgments without receiving feedback or explanations. After the testing phase, participants completed the satisfaction and trust measures before concluding the study. All the items presented in both phases were randomly re-ordered for each participant to control for possible order effects. On completion, participants were debriefed on the background of the study and paid £ 2.50 for taking part. Ethics approval for the study was granted by University College Dublin with the reference code *LS-E-20-11-Warren-Keane*. Task instructions and data for the study are at https://osf.io/smupq/.

3.2 Method: Participants

Participants (N = 207) were recruited using *Prolific Academic* (https://www.prolific.co), and assigned in a fixed order to three between-participant conditions: CF-Single (n = 68), CF-Group (n = 70) and CF-Group-Hint (n = 69). The sample consisted of 122 women, 84 men and one non-binary participant aged 18–76 years ($M = 40.86$, $SD = 14.31$), with respondents pre-screened to select native English speakers from Ireland, the United Kingdom, the United States, Australia, Canada and New Zealand, who had not participated in previous related studies. Prior to analysis, 19 participants were removed as they failed >1 attention or memory check. An *a priori* power analysis with G*Power indicated that 207 participants were required to achieve 90% power for a medium effect size with alpha <.05 for two-tailed tests.

3.3 Materials: Dataset and Implementations

The Adult (Census) dataset describes census information[2] and contains a mixture of continuous (age, weekly work hours, capital gain, capital loss), and categorical variables (employment type, occupation, marital status, country of birth, gender, race). A model that predicted a binary outcome, whether a person is earning over or under $50k/year, was implemented using a Gradient Boosting Classifier [13], as the to-be-explained black-box model. The default sklearn hyperparameters are used with a log loss implemented and a learning rate of 0.1 achieving an accuracy of 0.874 on the task. Training and test data were split and pre-processed according to Klaise *et al.* [25], where categorical features were encoded ordinally when possible. Instances were randomly selected from the dataset as query-items for the study. The instances selected were all correct classifications, with low predicted class confidence (within 0.15 of the other class) to make the classifications non-obvious to participants, and the selection was balanced with equal numbers for each class. For the training phase, 8 seed-instances were randomly selected (balanced across classes) and 4 nearest-like-neighbours were found for these seeds (using Hamming distance) to create the 40 queries used (i.e., eight 5-item sets of related instances). For the testing phase, all 40 items were randomly selected from the dataset as the queries to be used, balanced across classes. The single counterfactual explanations for each of the 40 training phase queries for the CF-Single condition were generated using DiCE [31][3]. For the CF-Group conditions, the Group-CF method (see Algorithm 1) was applied to the 5-item sets to find good group-counterfactuals to cover them. If a group-counterfactual was found then the five instances with their paired single- and group-counterfactuals were used as an item-set in the study. Finally, the sets of counterfactuals used in the CF-Single and CF-Group conditions were matched on proximity and sparsity. Proximity was measured using ℓ_1 distance scores, scaled by the median absolute deviation (MAD) of the feature's values in the training set (for continuous features), and a metric that assigns a distance of 1 if the features differ from the original input or zero otherwise (for categorical features). These distance scores were computed for the matched query-explanation pairs used in the control and experimental conditions; a paired two-tailed t-test indicated that they were not significantly different to one another, $t(39) = 1.30, p = .197$. Sparsity was measured as the number of feature-differences between counterfactual pairs which was always 2 for all query-explanation pairs used.

[2] The New Adult Datasets by Ding et al. [10] could be preferred here if examining group counterfactual explanations through the lens of fairness in XAI (see also [20]).

[3] We used the original random sampling variant of DiCE implemented using an sklearn backend and a sample size of 1000. The default post-hoc sparsity parameter of 0.1 and stopping threshold of 0.5 were implemented. Also, counterfactuals that we judged to be implausible were excluded from the materials.

3.4 Measures: Objective and Subjective Measures

A mix of objective and subjective measures was used (see [40] for a discussion of this distinction). The key objective measure was *accuracy*, which measured the extent to which exposure to the model's predictions and explanations in the training phase improved their knowledge of the domain/model; specifically, it was measured as the proportion of correct responses made in the training and testing phases (i.e., consistent with the model's predictions). The subjective measures evaluated people's self-assessments of their (i) *confidence* in their own predictions made in the testing phase, (ii) *satisfaction* with explanations used overall by the AI system, (iii) *trust* in the overall AI system. The latter two measures were made after people completed the training and testing phases of the study using the Explanation Satisfaction and Trust scales [15]. Four attention checks were deployed at randomised intervals (two during the training phase, and two during the testing phase) and participants were also asked to recall and select a subset of the features used by the system from a list of 10 options (5 correct, 5 incorrect) at the end of the experiment.

3.5 Results and Discussion

Figure 4 and Table 1 show the results across the three conditions in the study – CF-Single, CF-Group, and CF-Group-Hint – for the different measures used. On all the measures, the relative differences show a trend favoring the use of group counterfactuals but the conditions were not significantly different. Hence, as a first test of the idea, this experiment does *not* confirm that group counterfactuals are better than single counterfactuals in this task. Though this result may be initially disappointing to proponents of group-counterfactuals, it highlights the importance of performing human evaluations in XAI and incorporating

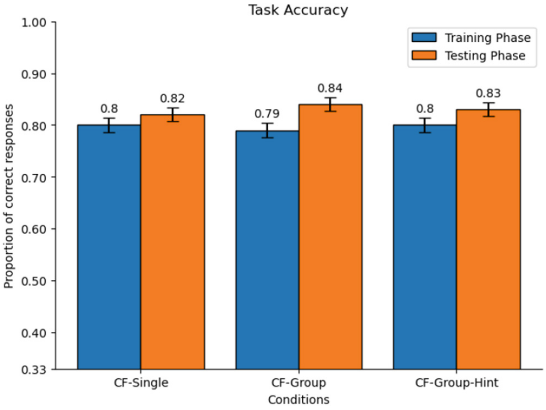


Fig. 4. Task Accuracy (proportion of correct answers) for the three conditions (CF-Single, CF-Group, CF-Group-Hint) in the Training and Testing Phases of the Study (showing standard error of the mean; y-axis begins at 0.33, chance-level for responding).

them in the development process of new methods, in order to ensure that explanations achieve their intended purpose. Furthermore, we explore the results in more detail to reveal some interesting insights into the conditions under which group counterfactuals might work. Specifically, this exploration shows that group counterfactuals may need to be presented (i) together as a consecutive group of explained instances rather than being randomly mixed up (as was done in the experiment), and/or (ii) in a task context where the group is explicitly assessed as a data pattern (e.g., as a risk analyst might do).

3.5.1 Objective Measure: Accuracy

A 3 (Explanation: CF-Single vs CF-Group vs CF-Group-Hint) x 2 (Phase: Training vs Testing) mixed ANOVA with repeated measures on the second factor was carried out on the proportion of correct responses given by each participant (i.e., accuracy; see Fig. 4). There was no main effect of Explanation $F(2, 204)=.174$, $p=.840$, however, there was a main effect of Phase, $F(1, 204)=25.153$, $p<0.001$, $\eta_p^2=.11$. As one would expect, participants in all three conditions were more accurate in the testing phase ($M=.829$, $SD=.096$) than in the training phase ($M=.795$, $SD=.111$). The two factors did not interact $F(2, 204)=1.608$, $p=.203$. Notably, response accuracy was quite high from the outset in the training phase (∼80% in all three conditions), potentially leading to ceiling effects in the testing phase. However, in the testing phase, there was still evidence of a trend in increasing accuracy across conditions with the following order: CF-Single < CF-Group < CF-Group-Hint conditions, Page's L(40)=500.0, $p=.013$ (see Fig. 4). This trend indicates that when people were given group counterfactual explanations (without and with a hint), their accuracy progressively improved. Indeed, the improvement in accuracy (the difference between a given participant's training accuracy and their testing accuracy), shows the CF-Single condition to be the least improved ($M=0.019$, $SD=.089$), with the CF-Group ($M=0.049$, $SD=.083$) and CF-Group-Hint conditions showing more improvement ($M=0.034$, $SD=.117$).

Table 1. Means and standard deviations for each measure in the conditions of the user study (CF-Single, CF-Group, CF-Group-Hint) for (i) *Accuracy* (proportion of correct answers in the testing phase), (ii) *Confidence* (ratings on a 5-point scale of each answered item in the testing phase), (iii) *Explanation Satisfaction* (summed ratings from the 8-item scale after the testing phase), (iv) *Trust* (summed ratings from the 8-item scale after the testing phase).

Measure	CF-Single		CF-Group		CF-Group-Hint	
	M	*SD*	*M*	*SD*	*M*	*SD*
Accuracy	0.816	0.104	0.839	0.085	0.832	0.100
Confidence	3.956	0.432	3.936	0.437	4.047	0.432
Satisfaction	26.632	6.462	25.800	6.826	28.493	6.910
Trust	24.324	5.454	24.786	5.522	25.899	5.480

3.5.2 Subjective Measures

A series of subjective evaluations were made by participants in the study, comprising confidence in their responses and satisfaction and trust in the AI system (see Table 1 for means and standard deviations). Overall, these measures show no main effects between conditions, though reliable trends are found in increasing scores across conditions with the following order: CF-Single < CF-Group < CF-Group-Hint. This trend was significant for *Confidence*, Page's L(40) = 513.5, p<.001, *Explanation Satisfaction*, Page's L(8) = 105.0, p = .012, and *Trust*, Page's L(8) = 108.0, p = .001.

3.5.3 Effects of Grouping on Accuracy

These weak effects could have emerged from the fact that the item-sets were not explicitly presented together in a group (recall, order was randomised). To address this question, we analysed the specific item-sets used in the training phase of the study. In the training phase, participants were presented with eight distinct 5-item-sets (i.e., 40 items in total); that is, sets of 5 similar instances and predictions that were explained with the same group counterfactual in the experimental conditions (CF-Group and CF-Group-Hint) and matched with respect to the control condition (CF-Single). For each participant, the 40 items were randomly re-ordered to control for possible order effects between item-sets. However, due to this randomisation, different participants would have seen more and less favourable sequences for a given item-set; that is, the 5 items in a set could happen to be presented together in the randomised sequence (with no gaps between them) whereas another item-set could be mixed in with other item-sets (with many gaps between items from the same set). Hence, some participants could have been presented with five (grouped) counterfactual explanations one after the other using the same target-feature values, presumably making it easier for them to benefit from the group-counterfactual. So, if a given item-set has lower gap-scores, one would expect higher accuracy for that set in both CF-group conditions. Accordingly, we analysed the order of items presented and calculated gap-scores for each item-set presented to each participant in the three conditions of the study to check whether favourable orderings had any effect. Spearman's correlations were computed between these gap-scores for item-sets in the training phase and the accuracy observed in that phase. This analysis showed that there were moderate-to-high, negative correlations between gap-scores and accuracy in all three conditions, the lower the gap-score between items from the same set, the higher the accuracy: CF-Single ($r_S(6) = -0.43$), CF-Group ($r_S(6) = -0.38$), and CF-Group-Hint ($r_S(6) = -0.69$). Notably, the CF-Group-Hint condition had the highest correlation, where participants were told that certain items were part of a group. Though these correlations do not imply causality, they present a consistent picture of the effects of group counterfactual explanations with a hint. It may be that instances need to be grouped in an unbroken sequence, to allow end-users to benefit from these group explanations.

4 Conclusion and Future Directions

This paper tested counterfactual explanations for multiple instances, or group counterfactuals, a new and emergent area of XAI. Proponents of this explanation strategy have argued that there are many real-world contexts in which various stakeholder users require predictions to be explained as meaningful groupings to provide additional insights (e.g., identifying disease patterns in a dairy herd, patterns of attrition in employees, risk profiles in mortgage holders). We advanced a case-based method for computing these explanations and carried out the first user tests of this group counterfactual concept. The results of this study showed weak support for the learning impacts of using group counterfactuals over single counterfactuals. However, group counterfactuals may work better in task contexts aimed at showing patterns in data (e.g., explicitly presenting pools of queries together). Specifically, the present study shows that the predictive-instances grouped in the explanation may need to be presented together to impact human learning. It is clear that more attention needs to be given to framing the task in which group counterfactuals are used. Our findings underscore the need for XAI research to move beyond traditional, simplified scenarios to more complex real-world, user-driven solutions.

Acknowledgements. This research was supported by Science Foundation Ireland via the Insight SFI Research Centre for Data Analytics (12/RC/2289-P2). For the purpose of Open Access, the author has applied a CC BY copyright to any Author Accepted Manuscript version arising from this submission.

References

1. Artelt, A., Gregoriades, A.: “how to make them stay?”–diverse counterfactual explanations of employee attrition. arXiv preprint arXiv:2303.04579 (2023)
2. Artelt, A., Gregoriades, A.: A two-stage algorithm for cost-efficient multi-instance counterfactual explanations. arXiv preprint arXiv:2403.01221 (2024)
3. Barocas, S., Selbst, A.D., Raghavan, M.: The hidden assumptions behind counterfactual explanations and principal reasons. In: Facct-20, pp. 80–89 (2020)
4. Byrne, R.M.: Counterfactuals in explainable artificial intelligence (XAI): evidence from human reasoning. In: IJCAI-19, pp. 6276–6282 (2019)
5. Carrizosa, E., Ramírez-Ayerbe, J., Morales, D.R.: Generating collective counterfactual explanations in score-based classification via mathematical optimization. Expert Syst. Appl. **238**, 121954 (2024)
6. Carrizosa, E., Ramírez-Ayerbe, J., Morales, D.R.: Mathematical optimization modelling for group counterfactual explanations. Eur. J. Oper. Res. (2024)
7. Celar, L., Byrne, R.M.: How people reason with counterfactual and causal explanations for artificial intelligence decisions in familiar and unfamiliar domains. Memory Cogn. **51**, 1481–1496 (2023)
8. Dai, X., Keane, M.T., Shalloo, L., Ruelle, E., Byrne, R.M.: Counterfactual explanations for prediction and diagnosis in XAI. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, pp. 215–226 (2022)
9. Dandl, S., Casalicchio, G., Bischl, B., Bothmann, L.: Interpretable regional descriptors: hyperbox-based local explanations. arXiv preprint arXiv:2305.02780 (2023)

10. Ding, F., Hardt, M., Miller, J., Schmidt, L.: Retiring adult: new datasets for fair machine learning. Adv. Neural Inf. Process. Syst. **34**, 6478–6490 (2021)
11. Dua, D., Graff, C.: UCI machine learning repository (2017)
12. Edwards, B.J., Williams, J.J., Gentner, D., Lombrozo, T.: Explanation recruits comparison in a category-learning task. Cognition **185**, 21–38 (2019)
13. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232 (2001)
14. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. (CSUR) **51**(5), 93 (2018). https://doi.org/10.1145/3236009
15. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for explainable ai: challenges and prospects. arXiv preprint arXiv:1812.04608 (2018)
16. Johnson, S.G., Johnston, A.M., Toig, A.E., Keil, F.C.: Explanatory scope informs causal strength inferences, pp. 2453–2458 (2014)
17. Kanamori, K., Takagi, T., Kobayashi, K., Ike, Y.: Counterfactual explanation trees: transparent and consistent actionable recourse with decision trees. In: AISTAT-22, pp. 1846–1870. PMLR (2022)
18. Karimi, A.H., Barthe, G., Schölkopf, B., Valera, I.: A survey of algorithmic recourse: contrastive explanations and consequential recommendations. ACM Comput. Surv. **55**(5), 1–29 (2022). https://doi.org/10.1145/3527848
19. Karimi, A.H., Schölkopf, B., Valera, I.: Algorithmic recourse: from counterfactual explanations to interventions. In: Facct-21, pp. 353–362 (2021)
20. Kasirzadeh, A., Smart, A.: The use and misuse of counterfactuals in ethical machine learning. In: Facct-21, pp. 228-236 (2021)
21. Keane, M.T., Kenny, E.M., Delaney, E., Smyth, B.: If only we had better counterfactual explanations: five key deficits to rectify in the evaluation of counterfactual xai techniques. In: IJCAI-21, pp. 4466–4474 (2021)
22. Keane, M.T., Smyth, B.: Good counterfactuals and where to find them: a case-based technique for generating counterfactuals for explainable AI (XAI). In: Watson, I., Weber, R. (eds.) ICCBR 2020. LNCS (LNAI), vol. 12311, pp. 163–178. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58342-2_11
23. Keil, F.C.: Explanation and understanding. Ann. Rev. Psychol. **57**, 227–254 (2006)
24. Kenny, E.M., Keane, M.T.: On generating plausible counterfactual and semi-factual explanations for deep learning. In: AAAI-21, vol. 35, no. 13, pp. 11575–11585 (2021)
25. Klaise, J., Van Looveren, A., Vacanti, G., Coca, A.: Alibi: algorithms for monitoring and explaining machine learning models (2020)
26. Kuhl, U., Artelt, A., Hammer, B.: Keep your friends close and your counterfactuals closer. In: Facct-22, pp. 2125–2137 (2022)
27. Langer, M., et al.: What do we want from explainable artificial intelligence (xai)?-a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research. Artif. Intell. **296**, 103473 (2021)
28. Lewis, D.: Counterfactuals. John Wiley & Sons, Hoboken (2013)
29. Lombrozo, T.: Explanatory preferences shape learning and inference. Trends Cogn. Sci. **20**(10), 748–759 (2016)
30. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2019)
31. Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: Facct-20, pp. 607–617 (2020)
32. Pakrashi, A., et al.: Early detection of subclinical mastitis in lactating dairy cows using cow level features. J. Dairy Sci. **106**, 4978–4990 (2023)

33. Plumb, G., Terhorst, J., Sankararaman, S., Talwalkar, A.: Explaining groups of points in low-dimensional representations. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, pp. 7762–7771 (2020)
34. Rawal, K., Lakkaraju, H.: Beyond individualized recourse: interpretable and interactive summaries of actionable recourses. Adv. Neural. Inf. Process. Syst. **33**, 12187–12198 (2020)
35. Read, S.J., Marcus-Newhall, A.: Explanatory coherence in social explanations: a parallel distributed processing account. J. Pers. Soc. Psychol. **65**(3), 429–447 (1993)
36. Ryan, C., Guéret, C., Berry, D., Corcoran, M., Keane, M.T., Mac Namee, B.: Predicting illness for a sustainable dairy agriculture: predicting and explaining the onset of mastitis in dairy cows. arXiv preprint arXiv:2101.02188 (2021)
37. Verma, S., Dickerson, J., Hines, K.: Counterfactual explanations for machine learning: a review. arXiv preprint arXiv:2010.10596 (2022)
38. van der Waa, J., Nieuwburg, E., Cremers, A., Neerincx, M.: Evaluating xai: a comparison of rule/example-based explanations. Artif. Intell. **291**, 103404 (2021)
39. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harvard J. Law Technol. **31**, 841 (2018)
40. Warren, G., Byrne, R.M.J., Keane, M.T.: Categorical and continuous features in counterfactual explanations of AI systems. In: IUI '23 (2023)
41. Warren, G., Keane, M.T., Gueret, C., Delaney, E.: If Only...If Only...If Only...we could explain everything. In: IJCAI-23 XAI Workshop (2023)