Actionable Recourse via GANs for Mobile Health

Jennifer Chien  
benshi.ai, UC San Diego

Anna Guitart  
Ana Fernández del Río  
África Periánez  
benshi.ai

Lauren Bellhouse  
Maternity Foundation

Abstract

Mobile health apps provide a unique means of collecting data that can be used to deliver adaptive interventions (Marsch, 2021b; Buckeridge, 2020; Overdijkink et al., 2018; Dolley, 2018; Dowell et al., 2016). The predicted outcomes considerably influence the selection of such interventions. Recourse via counterfactuals provides tangible mechanisms to modify user predictions. By identifying plausible actions that increase the likelihood of a desired prediction, stakeholders are afforded agency over their predictions. Furthermore, recourse mechanisms enable counterfactual reasoning that can help provide insights into candidates for causal interventional features. We demonstrate the feasibility of GAN-generated recourse for mobile health applications on ensemble-survival-analysis-based prediction of medium-term engagement in the Safe Delivery App, a digital training tool for skilled birth attendants.

Keywords: generative adversarial networks, recourse, counterfactual explanations, explainability, actionability, mobile health, survival analysis, user behavior

1. Introduction

The optimization of mobile health (mHealth) application ecosystems has attracted considerable interest in recent years. Many such applications are being designed to support front-line healthcare workers through capacity building, provision of clinical reference guidelines and diagnostic/triage support, realization of patient tracking and reporting, and more. Additionally, these apps can help healthcare providers overcome challenges in limited resource settings (Overdijkink et al., 2018; Hosny and Aerts, 2019b; Wahl et al., 2018b; Forero et al., 2018).

In particular, mHealth apps allow for personalized, just-in-time interventions (Marsch, 2021b; Buckeridge, 2020; Overdijkink et al., 2018; Dolley, 2018; Dowell et al., 2016). Such interventions can have far-reaching downstream effects, leading to better practices, and ultimately, better health outcomes.

This work examines the role of recourse via generative adversarial networks (GANs) (Goodfellow et al., 2014) in mHealth adaptive interventions. Recourse on predicted outcomes serves two roles: First, recourse identifies the plausible ways in which users can change their predictions, which can allow stakeholders not only to better understand model predictions, but also to have...
agency over them. Second, recourse, as a feedback mechanism and learning opportunity, can provide insights into actionable intervention candidates that can modify both predictions and outcomes.

We demonstrate this by using GANs to produce counterfactual estimates for survival-ensemble-based predictions of medium-term engagement with the Safe Delivery App (Foundation, 2022b) (hereafter, the App). The App is a digital training tool developed by Maternity Foundation (Foundation, 2022a), the University of Copenhagen, and the University of Southern Denmark, which contains evidence-based obstetric and newborn-related guidelines for supporting skilled birth attendants.

This paper is organized as follows: the remainder of this section describes the intervention framework for mHealth and defines recourse and counterfactuals. Moreover, we introduce the App, briefly review related work, and highlight the key contributions of this research. Section 2 describes the models used to demonstrate the GAN-based recourse approach. Section 3 defines the model specifications. The results and concluding remarks are presented in Sections 4 and 5, respectively.

1.1. Predictions and MHealth Apps

The bidirectional communication channel provided by data-centric mHealth apps can facilitate the collection of valuable user behavior and health outcomes information. These data can inform personalized interventions to increase the likelihood of a desired outcome (Hosny and Aerts, 2019a; Wahl et al., 2018a; O’Connor, 2018; Marsch, 2021a). As such, personalized predictions can be used to determine optimal timing and behavioral interventions for users.

In this work, we use conditional survival forests (CSFs) (Wright et al., 2017) to predict user lifetime, i.e. how long time users will use an app before stopping altogether, i.e. churning. We use churn prediction to identify users to target for push-notification-based interventions for re-engagement. Within the App specifically, these interventions can aim to prioritize essential training for short-lifetime predicted users.  

1.2. Recourse and Counterfactuals

Recourse mechanisms for behavioral predictions allow stakeholders (i.e., decision-makers or users) to identify concrete actions to modify their prediction and provide users agency over such predictions (Wallin, 1992).

In the context of mHealth apps, we define recourse as the set of feature changes that result in the desired predicted outcome. Recourse allows the stakeholders to better understand the decision-making process and implement tangible actions to modify the predictions. Recourse mechanisms rely on contrastive/counterfactual explanations (Byrne, 2019; Artelt and Hammer, 2019; Zhang et al., 2019; Klaise et al., 2021), which inform a user why and how a decision was made. Moreover, recourse can be generated directly from observational data, which is particularly useful in settings in which real-world experiments are prohibitively expensive, unethical, or infeasible.  

1.3. Related Work

The studies most closely related to the present research objective are Nemirovsky et al. (2020, 2021), which describe recourse via GANs in the context of image generation, hiring, diabetes, and recidivism. Other studies focused on recourse generation include

1. For additional information on the App, see Appendix A.
2. For extended discussion on recourse characteristics, see Appendix B.
Wachter et al. (2017); Ustun et al. (2019); Karimi et al. (2020).

In this work, recourse is studied in the context of survival-ensemble-based predictions of the time to the event of interest (here, churn). This approach to churn prediction has been described in Periáñez et al. (2016); Bertens et al. (2017); Kim et al. (2018); Chen et al. (2019); Olaniyi et al. (2022).

The App data have been previously analyzed in the contexts of content demand prediction (Guitart et al., 2021a), content recommendation (Guitart et al., 2021b), and engagement analysis (Olaniyi et al., 2022). Moreover, the App’s impact has been analyzed (Lund et al., 2016; Ohusola Oladeji and Oladeji, 2022).

Furthermore, a platform to deliver adaptive interventions in mHealth solutions has been established (Tang et al., 2021). Artificial intelligence and machine learning applications in healthcare have been summarized in Davenport and Kalakota (2019), and such applications in the global health context have been examined in Hosny and Aerts (2019a); Wahl et al. (2018a).

1.4. Our Contribution
To the best of our knowledge, this work represents the first attempt at applying a GAN-based recourse approach to survival analysis and predictions in the context of mHealth interventions. The results demonstrate the two primary impacts of providing recourse: identification of candidate interventions and resulting prediction modification within an mHealth intervention framework.

2. Methods and Models
We describe the methods and models used, specifically, CSFs and their use in churn prediction (Appendix C) and recourse via GANs and the counterfactual estimation it enables (Section 2.1).

2.1. GANs and Counterfactual Reasoning
GANs (Goodfellow et al., 2014) are composed of two artificial neural networks: a generator that produces realistic synthetic data, and a discriminator that differentiates between the generator output and real data. A GAN is trained through an adversarial min-max game in which the generator and discriminator are alternately trained. GANs can provide actionable, realistic, fast actions by a single feed-forward pass through a neural network.

GANs directly learn the data manifold from the training data, which is advantageous for recourse: First, the GAN output is realistic because the pre- and post-recourse action data are similar to the data manifold. Second, model owners can use GANs as a model of the data manifold to simulate interventional distributions to narrow the solution space for interventions and identify important actions.

CounteRGANs (Nemirovsky et al., 2020) generate counterfactual reasoning via feasible changes that result in a specified classification. A CounteRGAN is a specialized GAN composed of a generator $G$, discriminator $D$, and fixed classifier $C$. We define the features of user $i$ as $x_i$, with the corresponding predicted and true labels being $C(x_i)$ and $y(x_i) \in \{0,1\}$, respectively, denoting $y = 1$ as the desired outcome. $D$ is trained on data from the underlying true data distribution $p_{data}$. In contrast, $G$ is trained only on data $x_i \forall i \text{ s.t. } y(x_i) = 0$, i.e., for all users that churn. The output of $G(x_i) = a_i$ is the delta change that when added to the user’s original feature results in the desired predicted class, $C(x_i + a_i) = 1$.

3. Model Specifications and Data
We use data sampled from the App users from January 1, 2018 and October 1, 2021.
from countries with the maximum App use (India, Myanmar, Ethiopia...). Further information on feature construction is in Appendix D.

The model setup is shown in Figure 1. We examine GAN-generated recourse for users predicted to engage with the App for less than 90 days (period considered a marker of medium-term engagement). Further information on CSF and CounteRGAN training is in Appendix E.

We train CounteRGANs with survival forests involving 1, 5, and 20 trees, henceforth referred to as model\_{number of trees}. We train CounteRGAN on various sizes of CSFs to determine the influence of the model performance on recourse effectiveness. In addition, we generate counterfactuals with regularized gradient descent (RGD) implemented in Alibi, as described in (Klaise et al., 2021) with the 20-tree survival forest as a baseline.

4. Results

We describe the observed model performance and time requirements in Section 4.1, and Section 4.2 discusses the efficacy and cost of recourse. For concreteness, we provide an example of required feature changes to achieve successful recourse to a single user in Appendix G.

4.1. Model Performance and Time Requirements

As expected, model accuracy increases with the forest size for both the initial and post-recourse classifier. In addition, we measure mean clock time to compute the recourse (in seconds), finding that recourse generation via GANs is faster than that based on RGD, as it only requires a single feed-forward pass through the generator network to provide personalized recourse (instead of solving an optimization problem per user). Additional details are presented in the appendix H.

4.2. Efficacy and Cost of Recourse

In this section, we examine the efficacy and cost of recourse as an informational tool to the range of feasible and infeasible actions. In addition, we demonstrate how recourse can be used as an auditing mechanism for identifying subgroups for performance disparities, as recourse is dependent on model predictions.

To quantify the efficacy of recourse via CounteRGANs, we measure the percentage of users provided effective recourse \(C(x_i + a_i) = 1\) and denied recourse \(C(x_i + a_i) = 0\). We find that the model\_20 CounteRGAN provides the highest percentage of effective recourse with 29.3% compared to 15.6% with the model\_20 RGD. See additional details in Appendix I.

Differences in the efficacy of recourse can serve as a dataset- and model-specific auditing mechanism. For instance, Figure 2 shows the pre- and post-recourse user features, separated by the true outcome (\(y = 0\) and \(y = 1\) for the left and right columns, respectively) and recourse efficacy (effective and ineffective recourse in the top and bottom rows, respectively). For ease of visualization, we show the features reconstructed under the top two components using principal component analysis (Pearson, 1901).

Two observations can be made: First, across users with ineffective recourse (bottom row of Figure 2), the \(y = 1\) individuals are confined to a narrower range of principal component 2 than those with \(y = 0\). This finding can inform the range of infeasible recourse actions. Similarly, a comparison of the recourse efficacy (top and bottom rows in Figure 2) shows that the recourse efficacy may be higher with larger values of principal component 1. Although these results are
Figure 1: CounteRGAN Architecture for the App Data. Two neural networks are used: generator $G$ trained to output residuals, and discriminator $D$ trained to distinguish real and augmented data. Gray tiles labeled “generator input” represent data for users that churned in <90 days ($y(x_i) = 0$). The red tiles labeled “counterfactuals” represent post-recourse users that are predicted to use the App for $\geq 90$ days ($C(x_i + a_i) = 1$). The patchwork tiles represent residuals $a_i$. The gray and red stacked tiles labeled “real samples” represent the training data $y(x_i) = 1$.

dataset specific, they can highlight how recourse computation and analysis can clarify the domain of feasible actions.

Second, we observe overlap in the regions occupied by both effective and ineffective recourse. Although this result indicates that the recourse actions being provided are realistic and feasible, it may also indicate that a one-size-fits-all approach to providing recourse is likely insufficient. There may be user subgroups to which we cannot provide effective recourse via GANs. These subgroups may require additional incentives or indicate model performance disparities.

Thus, even in the case of ineffective recourse, stakeholders can still learn from group differences. Recourse can be used as an auditing mechanism aimed at answering the following questions: Can we provide users with recourse? Which groups of users cannot be provided with effective recourse? Is the model performance inferior for this subgroup? For this subgroup, does the GAN require better data or do we need to train a different model for recourse estimation?

We provide further analysis and discussion of the cost of recourse in Appendix J.

5. Summary and Conclusions

We develop a GAN-based approach to provide recourse on survival-forest-based predictions and demonstrate its application to obtain medium-term engagement predictions for users of the Safe Delivery App. The findings highlight the potential of obtaining fast, effective recourse via GANs in mHealth applications, its use as an auditing mechanism to locate disparities in recourse feasibility across model performance and user groups,
Figure 2: Scatterplot of Pre-Recourse (circles) and Post-Recourse (triangles) User Features. The original training data are transformed via principal component analysis for ease of visualization. Users are separated by predicted and true outcomes: $y = 0$ and $y = 1$ on the left and right columns, respectively; effective recourse ($C = 1$) and ineffective recourse ($C = 0$) are on the top and bottom halves, respectively.

and its role as an informational agent to audit misalignment between predicted and real-world feature importance.

The proposed approach provides a pipeline for identifying intervention-worthy features and causal recourse action candidates, drawing inferences directly and solely out of observations. This approach can provide all stakeholders with greater understanding and agency over model predictions (and hence, prediction mediated interventions) and mechanisms affecting the actual outcomes that are being predicted, enabling the definition of additional interventions.

In this context, recourse via GANs is likely to emerge as a key component of trustworthy adaptive interventions for mobile health toolkits.

Acknowledgments

The authors thank Archana Choudhary for reviewing the manuscript. This work was supported, in whole or in part, by the Bill & Melinda Gates Foundation INV-022480. Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 Generic License has already been assigned to the Author Accepted Manuscript version that might arise from this submission.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen,
Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levienberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org.

André Artelt and Barbara Hammer. Efficient computation of counterfactual explanations of lvq models. arXiv preprint arXiv:1908.00735, 2019.

Paul Bertens, Anna Guitart, and África Periáñez. Games and big data: A scalable multi-dimensional churn prediction model. In 2017 IEEE Conference on Computational Intelligence and Games (CIG), pages 33–36. IEEE, 2017.

David L Buckeridge. Precision, equity, and public health and epidemiology informatics—a scoping review. Yearbook of Medical Informatics, 29(01):226–230, 2020.

Ruth M. J. Byrne. Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, pages 6276–6282. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/876. URL https://doi.org/10.24963/ijcai.2019/876.

Pei Pei Chen, Anna Guitart, and África Periáñez. The Winning Solution to the IEEE CIG 2017 Game Data Mining Competition. Machine Learning Knowledge. Extraction, pages 1(1), 252–264, 2019.

Love S.B Clark T.G, Bradburn M.J and Altman D.G. Survival analysis part i: Basic concepts and first analyses. British Journal of Cancer, 89(2):232–238, 2003. doi: https://doi.org/10.1038/sj.bjc.6601118.

Thomas Davenport and Ravi Kalakota. The potential for artificial intelligence in healthcare. Future Healthcare Journal, 6(2):94–98, 2019. ISSN 2514-6645. doi: 10.7861/futurehosp.6-2-94. URL https://www.rcpjournals.org/content/6/2/94.

Shawn Dolley. Big data’s role in precision public health. Frontiers in public health, page 68, 2018.

Scott F Dowell, David Blazes, and Susan Desmond-Hellmann. Four steps to precision public health. Nature, 540(7632):189–191, 2016.

Thomas Fleming and D Lin. Survival analysis in clinical trials: Past developments and future directions. Biometrics, 56(4):971–983, 12 2000. doi: 10.1111/j.0006-341x.2000.0971.x.

Roberto Forero, Shizahr Nahidi, Josephine De Costa, Mohammed Mohsin, Gerry Fitzgerald, Nick Gibson, Sally McCarthy, and Patrick Aboagye-Sarfo. Application of four-dimension criteria to assess rigour of qualitative research in emergency medicine. BMC health services research, 18(1):1–11, 2018.

Stephane Fotso et al. PySurvival: Open source package for survival analysis mod-
Maternity Foundation. Maternity foundation. https://www.maternity.dk/, 2022a.

Maternity Foundation. Safe delivery app. https://www.maternity.dk/safe-delivery-app/, 2022b.

United Nations Children’s Fund. Child mortality estimation, u. levels & trends in child mortality: Report 2020: Estimates developed by the un inter-agency group for child mortality estimation. Technical report, United Nations Children’s Fund, 2020.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, Advances in Neural Information Processing Systems, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Amir-Hossein Karimi, Julius Von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. Advances in neural information processing systems, 33:265–277, 2020.

Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjarongk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. arXiv preprint arXiv:1907.09615, 2019.

Janis Klaise, Arnaud Van Looveren, Giovani Vacanti, and Alexandru Coca. Alibi explain: Algorithms for explaining machine learning models. J. Mach. Learn. Res., 22:181:1–181:7, 2021.
ers, Dan Hogan, Vicki Flenady, J Frederik Frøen, Zeshan U Qureshi, Claire Calderwood, Suhail Shiekh, Fiorella Bianchi Jassir, Danzhen You, Elizabeth M McClure, Matthews Mathai, and Simon Cousens. Stillbirths: rates, risk factors, and acceleration towards 2030. The Lancet, 2016.

Stine Lund, Ida Marie Boas, Tariku Bedesa, et al. Association between the safe delivery app and quality of care and perinatal survival in ethiopia: A randomized clinical trial. JAMA Pediatrics, 170(8):765–771, 2016. doi: 10.1001/jamapediatrics.2016.0687. URL https://jamanetwork.com/journals/jamapediatrics/fullarticle/2529144.

Lisa A Marsch. Digital health data-driven approaches to understand human behavior. Neuropsychopharmacology, 46(1):191–196, 2021a.

Lisa A Marsch. Digital health data-driven approaches to understand human behavior. Neuropsychopharmacology, 46(1):191–196, 2021b.

Daniel Nemirovsky, Nicolas Thiebaut, Ye Xu, and Abhishek Gupta. Countergan: generating realistic counterfactuals with residual generative adversarial nets. arXiv preprint arXiv:2009.05199, 2020.

Daniel Nemirovsky, Nicolas Thiebaut, Ye Xu, and Abhishek Gupta. Providing actionable feedback in hiring marketplaces using generative adversarial networks. Proceedings of the 14th ACM International Conference on Web Search and Data Mining, 2021.

Siobhan O’Connor. Big data and data science in health care: What nurses and midwives need to know. Journal of Clinical Nursing, 27(15-16):2921–2922, 2018. doi: https://doi.org/10.1111/jocn.14164.

Babaniyi Yusuf Olaniyi, Ana Fernández del Río, África Periáñez, and Lauren Bellhouse. User engagement in mobile health applications. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 4704–4712, 2022.

Meron Tessema Olusola Oladeji and Bibilola Oladeji. Strengthening quality of maternal and newborn care using catchment based clinical mentorship and safe delivery app: A case study from somali region of ethiopia. International Journal of Midwifery and Nursing Practice, 5(1):13–18, 2022.

Sanne B Overdijkink, Adeline V Velu, Ageeth N Rosman, Monique DM Van Beukering, Marjolein Kok, and Regine PM Steegers-Theunissen. The usability and effectiveness of mobile health technology–based lifestyle and medical intervention apps supporting health care during pregnancy: systematic review. JMIR mHealth and uHealth, 6(4):e8834, 2018.

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11):559–572, nov 1901. doi: 10.1080/14786440109462720.

África Periáñez, Alain Saas, Anna Guitart, and Colin Magne. Churn prediction in mobile social games: Towards a complete assessment using survival ensembles. In 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pages 564–573, 2016.

J. R. Quinlan. Induction of decision trees. Machine Learning, 1(1):81–106, mar 1986. doi: 10.1007/bf00116251.
Dexian Tang, Guillem Francès, and África Periáñez. A data-centric behavioral machine learning platform to reduce health inequalities. *arXiv preprint arXiv:2111.11203*, 2021.

UNFPA. Cost of ending preventable maternal deaths. Technical report, United Nations Population Fund, 2020.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, jan 2019. doi: 10.1145/3287560.3287566. URL https://doi.org/10.1145%2F3287560.3287566.

Arnaud Van Looveren, Janis Klaise, Giovanni Vacanti, Oliver Cobb, Ashley Scillitoe, Robert Samoilescu, and Alex Athorne. Alibi Detect: Algorithms for outlier, adversarial and drift detection, 10 2022. URL https://github.com/SeldonIO/alibi-detect.

Sandra Wachter, Brent Daniel Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Cybersecurity*, 2017.

Brian Wahl, Aline Cossy-Gantner, Stefan Germann, and Nina R. Schwalbe. Artificial intelligence (AI) and global health: How can AI contribute to health in resource-poor settings? *BMJ Global Health*, 3(4):e000798, 2018a.

Brian Wahl, Aline Cossy-Gantner, Stefan Germann, and Nina R Schwalbe. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? *BMJ global health*, 3(4): e000798, 2018b.

David E. Wallin. Legal recourse and the demand for auditing. *The Accounting Review*, 67(1):121–147, 1992. URL http://www.jstor.org/stable/248023.

WHO and UNICEF. Every newborn: an action plan to end preventable deaths. Technical report, World Health Organization, 2014.

Marvin N. Wright, Theresa Dankowski, and Andreas Ziegler. Unbiased split variable selection for random survival forests using maximally selected rank statistics. *Statistics in Medicine*, 36(8):1272–1284, 2017. doi: https://doi.org/10.1002/sim.7212.

Yujia Zhang, Kuangyan Song, Yiming Sun, Sarah Tan, and Madeleine Udell. “why should you trust my explanation?” understanding uncertainty in lime explanations. *arXiv preprint arXiv:1904.12991*, 2019.

Appendix A. The Safe Delivery App

Maternal and neonatal mortality remains a pressing problem in global health. Each year, nearly 300,000 women and 5 million newborns die of causes directly related to pregnancy and childbirth (Fund, 2020; UNFPA, 2020). In addition, nearly all newborn deaths occur in low- and middle-income countries, 80% of which are preventable and treatable by cost-efficient interventions (WHO and UNICEF, 2014; Lawn et al., 2016).

Digital tools such as the App can help build the capacities of skilled birth attendants (Lund et al., 2016; Olusola Oladeji and Oladeji, 2022). The App is divided into clinical modules covering key topics related to maternal and newborn health, including both normal labour and birth as well as guiding healthcare workers through common obstetric and newborn complications. Each module consists of educational videos, easily
referenceable action cards, drug lists, practical procedures, and a series of tests that increase in difficulty to assess the depth of knowledge and skills acquired by users.

**Appendix B. Recourse Characteristics**

As defined, recourse must have the following four characteristics: First, recourse actions must be *actionable*, i.e., they must use features that users can change (Joshi et al., 2019; Karimi et al., 2020; Ustun et al., 2019). For example, recourse may be provided to a user for increasing their lifetime in the app, but not by changing the date of first use. Second, recourse actions must be *realistic* (Wachter et al., 2017), following real-world directional restrictions of features (e.g., preservation of the monotonically increasing nature of accumulated metrics). Third, recourse actions must be *feasible* (Ustun et al., 2019) and not prohibitively expensive to execute (e.g., increasing the number of modules completed weekly by a *reasonable* amount). Finally, recourse actions must *obtain the desired predicted outcome*, meaning that the implementation of such actions must result in the desired predicted outcome.

**Appendix C. CSFs and Churn Prediction**

Survival analysis models have historically been used in medical and biological research for estimating life expectancy (Clark T.G and D.G, 2003; Fleming and Lin, 2000; Quinlan, 1986). However, the survival analysis toolbox provides methods generalizable to modeling and predicting the time to an event of interest in the presence of right censoring. CSFs (Wright et al., 2017) are a variant of the random survival forests (RSFs) (Ishwaran et al., 2008) that recursively partition the feature space pertaining to the splitting criteria through linear rank statistics. This method minimizes the bias typically encountered in RSF predictions.

In this context, CSFs can be used to predict how long users will remain active before they churn (i.e., the number of days between the first and last login) (Periáñez et al., 2016; Bertens et al., 2017; Kim et al., 2018; Chen et al., 2019; Olaniyi et al., 2022). We define churn as a number of consecutive days without in-App activity, that is determined following Olaniyi et al. (2022).

**Appendix D. App Data and Feature Construction**

We use the data pertaining to a sample of users of the App between January 1, 2018 and October 1, 2021, from countries with the maximum App use (India, Myanmar, Ethiopia...). We use a train/test split of 50/50, resulting in datasets of 730 and 729 users, respectively. The in-App logs are processed into daily user metrics, such as the number of daily sessions, time spent using the App, days since last login, and similar values for specific e-learning contents to characterize the user activity and behavior. The metrics are used to construct features by applying statistical operations (mean, max, normalization, etc.) over different periods, resulting in 181 features.

**Appendix E. Model Training**

First, the CSF models are trained and used to predict days to churn for the users. Next, the predicted days to churn are converted to a binary label predicting whether the user will have a lifetime of at least 90 days. Next, we input feature data from users that churned in less than 90 days to $G$ and train it iteratively via gradient descent. We add the generator output and original input to construct the counterfactuals of users...
that achieve a predicted lifetime of at least 90 days and appear realistic to the discriminator. The discriminator is trained to distinguish users that will churn in 90 days or more and those that have implemented recourse actions to extend their predicted lifetime to at least 90 days.

Each CounteRGAN is trained for up to 600 iterations but saved at checkpoints at which sufficient discriminator performance is observed. These checkpoints correspond to an accuracy less than or equal to 0.55 for the pre- and post-recourse users. Low accuracy on pre- and post-recourse data indicate that the post-recourse data are difficult to distinguish and thus more realistic.

Appendix F. Metric Definitions

The CSF and GAN performances are evaluated using the following metrics:

- **Accuracy:** \( \frac{1}{n} \sum_i^n I(y_i, \hat{y}_i) \).
  - Model Accuracy: \( n \) is the total number of users in a dataset (training or testing, \( y \in \{0, 1\} \) or \( y \in \{0\} \), respectively). The accuracy indicates the model performance in terms of the number of predictions that match the true observed label.
  - Discriminator Accuracy: This metric indicates the discriminator performance in terms of the number of predictions that match the true original data or data augmented by the generator.
  - Classifier Accuracy: \( n \) is the total number of \( y = 0 \) users that are augmented sufficiently by the generator to obtain a \( \hat{y} = 1 \) prediction.

- **Percent Denied:** \( \frac{1}{n} \sum_i^n I(\hat{y}(x_i), 0) \), where \( n \) is the number of applicants in the dataset (training/testing). The metric indicates the proportion of applicants predicted to churn in less than 90 days, thereby qualifying for recourse.

- **Percent Successful Recourse:**
  \[
  \frac{1}{\|\hat{y}(x_i + a_i) = 0\|} \sum_{i \in \{\hat{y}(x_i + a_i) = 0\}} I(\hat{y}(x_i + a_i), 1).
  \]
  This metric indicates the proportion of applicants predicted to churn more than 90 days after adopting the recourse actions recommended by the generator.

- **Mean Cost of Successful Recourse Actions:**
  \[
  \frac{1}{\|\hat{y}(x_i + a_i) = 1\|} \sum_{i \in \{\hat{y}(x_i + a_i) = 1\}} ||a_i||^2.
  \]
  This metric indicates the average \( L_2 \) distance for the feature change for applicants who successfully achieve the desired predicted outcome.

- **Cumulative Cost of Denied Recourse:**
  \[
  \sum_{i \in \{\hat{y}(x_i + a_i) = 0\}} ||a_i||^2.
  \]
  This metric indicates the total \( L_2 \) distance for the feature change for applicants who do not achieve their predicted desired outcome.

- **Average Clock Time:**
  \[
  \frac{1}{n} \sum_i^n T_{end}(x_i) - T_{start}(x_i),
  \]
  where \( n \) is the number of users in the dataset (training/testing), and \( T \) is the clock time at which the recourse calculation starts and ends. This metric indicates the average time (in seconds) required to calculate a user’s recourse actions to estimate the corresponding counterfactual.

Appendix G. Recourse Action Set Example

Appendix H. Model Performance and Time Requirements

The initial classifier performance over all users and those who churned in less than 90 days is presented in Table 2. As expected, the model performance is enhanced with the forest size, from 0.7177 for all users
Table 1: Example of a recourse action. We provide the top five largest changes constituting a successful recourse action for a single user under model_20. “Original” indicates the user’s original feature values, and “Required” lists the feature values required for the users to be predicted to churn in ≥ 90 days. All features here are maximum normalized. In this case, the ‘norm_max’ extension on a feature indicates the maximum normalized

for model_1 to 0.8709 for model_20. The percent of denied users is consistent across models (Table 3), increasing slightly from 44.96% for model_1 to 52.34% for model_20. The lack of significant variation is attributable to the binarization of the survival prediction.

Low discriminator performance on both real and fake data signifies realistic “generated” (transformed) data (Table 2). In addition, for post-recourse applicants, the model performance is enhanced with the forest size. This phenomenon likely occurs because higher accuracy CSFs can provide more precise feedback to the GAN, allowing it to better learn the data manifold and differentiate classes. Therefore, model_20 GANs represent the optimal platform.

Appendix I. Efficacy of Recourse
Table 3 summarizes the values of the percentage of users provided $C(x_i + a_i) = 1$ and denied recourse $C(x_i + a_i) = 0$ after implementing a CounteRGAN-generated action. Although the percentage of recourse-denied users is consistent across models, the percentage afforded recourse increases for larger forests. This result supports that CounteRGANs trained on more accurate models may be able to provide more effective recourse. In addition, both models_5 and _20 provide more effective recourse than that of counterfactuals obtained via RGD.

Appendix J. Cost of Recourse
We examine the cost of recourse as another auditing mechanism. We estimate the cost of recourse with $L_2$ distance across effective and ineffective recourse. Table 3 presents the mean cost of an effective recourse action and cumulative denied recourse cost across models. Model_1 has the lowest mean cost of successful recourse actions, followed by models_20 and _5. However, the cumulative denied recourse cost does not exhibit the same trend: it increases from model_1 to models_5 and _20, thereby highlighting the differences
Table 2: Validation Metrics. Accuracy metrics for the proposed counterfactual models based on GANs with CSFs with varying number of trees 1, 5, 20 converted to binary classifiers. Results of a baseline model that produces counterfactuals via regularized gradient descent with a 20-tree CSF converted to a binary classifier are also presented for comparison.

| Counterfactual Model | Conditional Survival Forest Num Trees | Initial Model Accuracy \((y \in \{0, 1\}/\{0\})\) | Discriminator Accuracy \((D(x_i)/D(x_i + a_i))\) | Post-Recourse Classifier Accuracy |
|---------------------|----------------------------------------|-------------------------------------|-----------------------------|-----------------------------|
| via GANs            | 1                                      | 0.718 / 0.656                      | 0.194 / 0.806              | 0.372                      |
|                     | 5                                      | 0.813 / 0.810                      | 0.613 / 0.278              | 0.497                      |
|                     | 20                                     | 0.871 / 0.868                      | 0.485 / 0.512              | 0.627                      |
| via RGD             | 20                                     | 0.871 / 0.868                      | -                          | 0.627                      |

Table 3: Recourse Generation Metrics. Metrics to evaluate the recourse performance, cost, efficiency and runtime for the proposed counterfactual models and baseline counterfactual model. Results of a baseline model that produces counterfactuals via regularized gradient descent with a 20-tree CSF converted to a binary classifier are also presented for comparison.

| Counterfactual Model | Conditional Survival Forest Num Trees | % Denied \((C = 0)\) | % Successful Recourse \((C = 1)\) | Mean Cost of Successful Recourse Actions | Cumulative Cost of Denied Recourse | Mean Compute Time (seconds) |
|---------------------|----------------------------------------|---------------------|---------------------------------|------------------------------------------|---------------------------------|---------------------------|
| via GANs            | 1                                      | 45.0%               | 4.7%                            | 1.5496                                   | 319.22                          | 0.0006                    |
|                     | 5                                      | 51.9%               | 16.1%                           | 1.7873                                   | 441.45                          | 0.0003                    |
|                     | 20                                     | 52.3%               | 29.3%                           | 1.6833                                   | 535.30                          | 0.0005                    |
| via RGD             | 20                                     | 52.3%               | 15.6%                           | 0.0702                                   | 3.45                            | 132.3800                  |

in the cost of recourse across user groups. Figure 3 shows the differences in the cost distributions of effective (left) and ineffective (right) recourse. The cost of effective recourse is similar across models. However, the cost of ineffective recourse differs across models, with model 20 associated with a wider range but lower frequency.

This finding demonstrates that recourse may be providing disproportionate gains and losses across those who received effective and ineffective recourse. In addition, the comparison of recourse via CounteRGANs and RGDs shows that RGDs have a lower cost, in terms of both the mean cost of effective recourse action and cumulative denied cost. The effects of lower efficacy recourse must be considered as it may have negative downstream effects such as lack of accountability, lack of trust, and potential disengagement with the App.

To demonstrate the utility of recourse as an informational tool, stakeholders can examine the distributions of the cost of recourse across true and predicted outcomes. Figure 4 shows two histograms of the cost
Figure 3: Histogram of the Recourse Cost by Efficacy. Cost distribution for the different models for effective (left) and ineffective (right) recourse. For the considered dataset, model_20 is associated with a wider range of recourse cost than model_5 and model_1.

of recourse associated with a particular feature: the maximum number of actions taken in the last 15 days. The plot to the left shows the distribution across effective and ineffective recourse actions. The plot to the right shows the corresponding distribution across true outcomes.

The cost distributions differ when separated by the predicted outcome and true label. This finding indicates that this feature may not be causally related to the true outcome and may thus have an inflated estimation effect on the true outcome. Examining these distributional differences across important features can help audit the model trust, especially as it is used to provide recourse suggestions to users.

Appendix K. Data and Code Availability

All data used in this analysis are derived from the Safe Delivery App logs and belong to the Maternity Foundation. For inquiries regarding the use of these data, please contact the Maternity Foundation at mail@maternity.dk. The code used is available at https://github.com/benshi-ai/paper_2022_ml4h_recourse_gans.

All the analysis were performed with Python version 3.7.13 on Google Colab, using the following packages from PyPI: pysurvival 0.1.2 (Fotso et al., 2019–), alibi 0.7.0 (Van Looveren et al., 2022), and tensorflow 2.8.2 (Abadi et al., 2015).
Figure 4: Histograms of the Recourse Cost Associated With a Specific Feature Separated by Recourse Efficacy (left) and True Outcome (right). The feature is the maximum normalized number of actions taken by a user in a day over the last 15 days that the user was active. The costs correspond to model 20. The distributions across the predicted and true outcome differ in terms of the range, i.e., $y = 0$ has a wider range than $\hat{y} = 0$, indicating disparities in the estimated and true causal impacts of a feature.