ETIC 2.0 設計・検証ノート07

その人に役立つAI支援を、どう確かめるか

個人内の比較で分かること、分からないこと



Daiki Morimoto / Unknown Lab
ORCID: 0009-0009-6444-750X
Version 1.0 / 公開日 2026-10-18
DOI: 10.5281/zenodo.23447401
未査読 / 著者確認済み
独立専門レビュー未実施 / AI支援執筆
CC BY 4.0 / Copyright 2026 Daiki Morimoto

30秒で読む

AIが本人の生活に合わせて提案を変えていても、その提案が本人に役立つと確かめられたことにはならない。出かけやすい日に多く提案する仕組みなら、提案後の外出が多くても、提案のおかげか、もともと出かけやすかったのかを分けられない。

本稿の答えは、本人が変えたいことを一つ定め、安全に変更できる支援条件を、同じ人の比較可能な機会で繰り返し比べることである。 ただし、効果が後まで残る支援や、途中で中身が変わるAIには、単純な切替比較をそのまま使えない。比較できない条件を残すことも、検証の一部になる。

前稿は、AIがケアの相談や連絡をどこまで引き受けられるかを扱った。今回は、集団での性能や平均的な有用性から一歩進め、目の前の本人について、支援を続ける、変更する、判断を保留する根拠を考える。N-of-1試験などの既存の研究方法を用いた検討であり、新しい効果を実証した報告ではない。

1. 「あなた向け」と「あなたに効く」の間

生活支援アプリが、予定や過去の利用履歴に応じて「そろそろ出かける準備を」と知らせる。毎日同じ時刻に届く通知より、自分の事情を分かってくれているように感じられる。これは使いやすさとして大切な経験である。しかし、本人に合わせて内容を変える仕組みと、その変更によって生活がよくなることは、別々に確かめる必要がある。

この違いを、架空の例で考える。図書館へ行く準備を後回しにしがちな人が、固定時刻の通知から、予定の空きを見て通知するAI機能へ切り替えた。切替後は外出が増えたため、「AIが自分に合う時刻を見つけた」と感じた、とする。

ところが、切替前は仕事の繁忙期で、切替後には休暇があった。AIは予定に余裕のある日に通知し、外出の記録もその日に集まっていた。この設定では、外出が増えた理由はAIの通知かもしれないし、休暇かもしれない。

通知された日だけを集計すると、通知に向く日をAIが選んだことまで、通知の効果に見えてしまう。

単に「もっと記録する」だけでは、この問題は解けない。同じ選び方を続けて記録を増やしても、通知と生活条件が一緒に変わるからである。ここで必要なのは、記録の量より、何と何を比べた記録なのかを明らかにすることである。

2. 同じ人の中で比べるとは何か

一人の中で複数の支援条件を計画的に切り替え、繰り返し評価する方法に、N-of-1試験がある。ここでNは対象となる人数を表す。一人を一度だけ観察するのではなく、その人にとっての比較を複数回つくる考え方である。英国の公的ガイダンスは、無作為に定めた順序で条件を経験する試験を説明し、時間の近い観測どうしの関連にも配慮した解析を求めている。[1]

ただし、「前半は通常機能、後半はAI機能」と分けるだけなら、時期の違いが残る。月曜は通常機能、日曜はAI機能と固定しても、曜日の違いと区別できない。どちらも同じ人の記録ではあるが、それだけで公平な比較になったわけではない。

先ほどの例なら、本人が図書館へ行こうと決め、外出可能と判断した機会を先に定義する。その範囲で、Aを固定時刻の通知、Bを予定に合わせた通知とし、専門家と設計した割付に従って比較する案が考えられる。通知内容や利用できる情報などはそろえ、主に比べたい違いを通知時刻の決め方に限定する。これは未実施の設計例であり、一定回数行えば効果が分かるという実施手順ではない。

AIに「成功しそうな日はB、難しそうな日はA」を選ばせれば、再び条件が偏る。どの機会にどちらを使うかを決める仕組みと、Bの通知時刻を決める仕組みを分ける必要がある。無作為化も、一人の短い期間で曜日や体調を完全に均等にする保証ではない。実際にどの条件が重なったかを確かめ、観測数と解析方法を事前に検討する。

比較の目的は「AIを使った回数」を増やすことではない。この人なら、決めていた図書館への外出を実行できたか、そのために余計な確認や催促が増えなかったかが重要になる。本人が予定を変えた日は、理由を確かめず失敗に数えない。機器が記録を落とした日も、外出しなかった日と区別する。

3. 通知研究が教える、比較できる範囲

日常の小さな支援を繰り返し比べる研究例がHeartStepsである。Klasnjaらは、状況に合わせた活動提案について、利用可能な時点ごとに提案の有無などを無作為に割り付けるマイクロランダム化試験を行った。分析対象は37人、期間は6週間で、評価したのは割付後30分間の歩数である。[2]

提案全体では、提案なしに対する歩数の比は1.14、95%信頼区間は0.99〜1.31、p値は0.06だった。平均14%増という推定値だけを取り出して、有効だったと断定できない。副次的に分けた歩行提案では増加が示されたが、効果は時間とともに弱まった。[2]

この研究は、通知の直後に何が変わったかを、日常の中で比較する実例である。一方、参加者のデータをまとめた推定は、そのまま37人全員に効果があった証明にはならない。また、歩数の記録がない時間をゼロと扱うことによる限界があり、一日の活動全体や健康改善を確かめた試験でもない。[2]

さらに、このHeartStepsは事前に用意された文面を状況に合わせて届ける仕組みであり、現在の生成AIの有効性を示してはいない。本稿が借りるのは、古いシステムの成績を新しいAIへ移す論法ではなく、支援を出す時点に比較を組み込み、何への効果を測ったかを限定する考え方である。

マイクロランダム化試験とN-of-1試験も同義ではない。前者で多数の人の時点データを集めても、主な問いが集団平均の短期効果なら、個々人の継続判断には追加の検討が要る。繰り返し観測したことと、個人ごとの判断に足る精度があることを分ける必要がある。

4. 比較した後、何を選ぶのか

架空例へ戻る。本人は、休暇と通知機能が同時に変わった最初の記録だけでは、Bが優れているとは結論しない。まず、必要な連絡や予定表はそのまま残し、変更してよい任意の通知だけを比較対象とする。本人が望む外出、通知の煩わしさ、記録の手間を、開始前に確認する。

比較が可能と判断された場合、結末は「AIの方がよかった」の一通りではない。ここでは結果を捏造せず、結果に応じて何を変えるかを先に示す。

比較で分かったこと	この例での判断案
Bで本人に意味のある改善があり、負担も許容できる	調べた生活条件の範囲でBの継続候補とする
改善の幅を絞れず、Bの手間だけが増える	優位性は保留し、本人が選ぶ負担の少ない方法へ戻す
Bは外出を増やすが、通知が苦痛になる	外出回数だけで採用せず、頻度や通知方法を変更する
途中で生活やAIの中身が大きく変わった	同一条件の結果としてまとめず、比較の前提を見直す

表1。本稿の架空例に対する未検証の判断案。統計的な採否基準や臨床上の閾値を定めた表ではない。

この例で確実に変わるのは、AIの採用を正当化する説明である。「自分向けだからよい」から、「どの条件と比べ、本人の何が変わり、何がまだ不確かか」へ変わる。差がはっきりしない場合にAを選んでも、Bに効果がないと証明したことにはならない。効果についての結論と、現在の負担を踏まえた生活上の選択は一致しないことがある。

何を意味のある違いと考えるかは、研究者のp値だけでは決まらない。DIAMONDの方法論研究は、本人にとって重要な結果、臨床的な意味、本

人・市民の関与を設計と解釈に位置づけている。[3] 比較の説明が正しくても、本人が望まない試験を続ける理由にはならない。

5. AIが学び続ける場合は、何を比べているのか

AIを含む支援では、利用中に仕組み自体が変わることがある。Bを使った期間の反応から通知規則が更新され、次のBでは別の判断をするなら、Bというラベルだけでは支援の中身を特定できない。Aへ戻した後にも、その学習結果が使われる設計なら、以前のBの影響がAへ持ち越される。

その場合には、二つの問いを分ける必要がある。固定した通知規則BがAよりよいのかを調べたいなら、比較中に何を固定できるかが問題になる。学習し続ける仕組み全体がよいのかを調べたいなら、学習の規則や履歴の扱いまでを介入として定義し、単純な切替とは別の設計・解析を検討する。モデルの更新日を書くだけで、この問題が解けるわけではない。

本人や支援者が変更できる範囲にも限界がある。通知の頻度や時刻を設定できても、提供者のモデルや学習データまで書き換えられるとは限らない。変更権限も内部記録もない部分について、「本人に合わせてチューニングして検証する」と一括して約束すべきではない。設定で変えられる部分、提供者の協力が必要な部分、現時点では観察しかできない部分を分ける。

この区別は、AIが高性能になれば不要になるものではない。予測が当たることと、別の支援を選んだ場合に本人がどうなったかを確かめることは異なる。AIが作った「もし別の方法だったら」という説明も、観測していない結果そのものではない。比較を助ける道具と、比較の代わりになる証拠を取り違えないことが重要である。

6. 元へ戻せない支援は、無理に切り替えない

一度学んだ技能や、新しく築いた人との関係は、AIを止めても消えない。支援によって外出の習慣が身についた場合にも、通知だけを切り替えて、最初と同じ状態へ戻ったと見なすことはできない。これらは前の支援の影響が後の期間に残る、持ち越しの問題である。

公的ガイダンスは、効果の立ち上がりが遅い場合、影響が長く残る場合、状態が急速に変わる場合に、単純なN-of-1試験が適さないことを示している。[1] 通知という軽い介入でも、習慣化が起きればこの制約を免れない。

だからといって、覚えたことを忘れさせたり、必要な支援を取り上げたりして比較条件をつくるわけではない。DIAMONDも、支援を中断することによる害に注意を促している。[3] 長期の学習や回復を評価するなら、その変化を含めた別の研究計画を方法論の専門家と検討する必要がある。

そもそも、本人が比較に意味を感じず、記録が負担になるなら、試験をしない選択がある。日々の観察や本人の選好を使って支援を決めることはできる。その際には、好みに基づく選択、経過の観察、因果効果を比較した判断を区別して記録すればよい。あらゆる生活上の選択に、個人実験を義務づける提案ではない。

7. よく測れば、生活も必ずよくなるのか

評価方法の精密さと、その方法を取り入れる便益も別問題である。Kravitzらは、慢性筋骨格痛のある215人を、スマートフォン等で支援する個別のN-of-1試験への参加群108人と、通常ケア群107人へ無作為に分けた。[4]

主要評価だった6か月後の痛みによる生活への支障の変化について、群間差は−1.36点、95%信頼区間は−2.91〜0.19点、p値は0.09だった。試験への参加が通常ケアを上回る改善をもたらしたとは結論できない。これは「両者は同等」と証明した結果でもない。[4]

この研究は生成AIの試験ではなく、図書館の架空例とも対象が異なる。それでも、本人向けに比較する方法を導入したこと自体から、健康改善を約束できないという反証的な手掛かりになる。正確な情報が得られても、選べる支援が変わらない、実行に負担がある、といった可能性は残る。後半は本稿の推論であり、この試験で改善が明確でなかった原因を特定した説明ではない。

AIが記録や集計を支援することは、検討に値する。しかし、記録時間が減った、判断が変わった、本人の生活が改善した、という三つを同じ成果として扱わない。測定を容易にしたAIの価値も、別途確かめる対象になる。

8. ETIC 2.0との関係

考え方と重なる点

ETIC 2.0は、技術を本人の生活や環境との関係で考える枠組みである。[5] 今回扱った方法は、集団の平均だけでなく、本人にとっての違いを検討する点で重なる。ただし、N-of-1試験やマイクロランダム化試験は既存の方法であり、ETIC 2.0の発明でも有効性の証明でもない。

答えるべき問い

ETIC 2.0で支援を見直す際、本人に合わせたという説明を超えて、何が変われば継続・変更の判断を変えるのかを示せるか。集団での知見、本人の観察、個人内の比較を区別し、比較できない部分を残せるか。特にAIが更新される場合、どの版と運用条件についての判断だったのかを追えるかが問われる。

具体的な見直し案

Unknown Labの未検証な提案として、既存の支援記録に「今回決めたいこと」「変更できる条件と比較相手」「本人に意味のある違い」「判断を保留・中止する条件」を結びつける。新しい尺度や総合点を増やすのではなく、既存のN-of-1設計や通常の経過記録で十分なら、それを用いる。

この案の追加価値は未検証である。整理項目を加えても判断が変わらず記録の手間だけが増えるなら、項目を削る。比較対象を定められないAI機能には、実証済みという表示を付けない。評価の厳密さを理由に、本人の選択や必要な支援を狭めない。

結論

この人に役立つAI支援を確かめるには、本人向けの提案を出すだけでは足りない。安全に変えられる条件について、比較相手と本人に意味のある結果を定め、不確かさと負担を含めて継続・変更を判断する必要がある。条件が元へ戻らない場合や、AIの中身を特定できない場合には、その比較から言えることを狭める。

本稿が前稿に加えるのは、AIへ何を任せるかという設計から、誰について、どの条件で、何が分かれば任せ方を変えるのかという検証への一歩である。AIが本人を理解しているように見えることよりも、本人の生活で確かめられた範囲を説明できることを重視する。

調査範囲と限界

2026年10月18日に行った、問いを限定したナラティブな検討であり、系統的レビューではない。デジタル支援の個人内比較を説明する公的ガイダンス、方法論研究、通知の原著試験、評価方法自体を比較した原著試験を選定した。2024〜2026年の更新も探索したが、新しい生成AIの個人別効果を実証したとは扱わない。

[1]の本文、[2]の方法・結果・限界、[3]の本文と訂正告知を確認した。[4]はPubMed掲載の抄録・図説明の確認に限定し、主要結果以外の詳細な解析や補足資料の結論は引用していない。原データの再解析と独立専門レビューは未実施である。架空例、AIへの適用上の整理、ETIC 2.0の見直し案は本稿の推論・提案であり、実施済み研究ではない。

著者・利益相反・位置づけ

Daiki Morimoto / Unknown Lab。ETIC 2.0の提案者による未査読の研究ノートであり、自説に関する知的な利害を有する。AI支援による調査・執筆・監査を含む。独立した方法論・専門職・当事者レビューは未了である。個別の診療判断や、服薬・必要なケアを停止する実施指示ではない。

参考文献

1. Office for Health Improvement and Disparities. N-of-1 study: comparative studies. GOV.UK. 2020-07-16. https://www.gov.uk/guidance/n-of-1-study-comparative-studies

2. Klasnja P, Smith S, Seewald NJ, et al. Efficacy of Contextually Tailored Suggestions for Physical Activity: A Micro-randomized Optimization Trial of HeartSteps. Annals of Behavioral Medicine. 2019;53(6):573–582. https://doi.org/10.1093/abm/kay067

3. Chatters R, Hawksworth O, Julious S, Cook A. The development of a set of key points to aid clinicians and researchers in designing and conducting n-of-1 trials. Trials. 2024;25:473. https://doi.org/10.1186/s13063-024-08261-z（訂正: 2024;25:600、補足資料2の文献表記。https://doi.org/10.1186/s13063-024-08453-7 ）

4. Kravitz RL, Schmid CH, Marois M, et al. Effect of Mobile Device-Supported Single-Patient Multi-crossover Trials on Treatment of Chronic Musculoskeletal Pain: A Randomized Clinical Trial. JAMA Internal Medicine. 2018;178(10):1368–1377. https://doi.org/10.1001/jamainternmed.2018.3981

5. Morimoto D. エコテクノロジカル・インテグラティブ・ケア(ETIC)2.0, Version 2.1改訂版: AI時代の作業療法とケア生態系設計. Unknown Lab; 2026. https://doi.org/10.5281/zenodo.21848456