Color and Emotion in User Experience Design for Older Adults' Health Applications: A Multi-Database Bibliometric and Evidence-Mapping Study

YangFan He[1], Azizah Che Omar[2]*, Sobihatun Nur Abdul Salam[3]
School of Multimedia Technology and Communication, Universiti Utara Malaysia, Sintok, Kedah, Malaysia[1, 2, 3]
Fujian Forestry Vocational & Technical College, Nanping, Fujian, China[1]

***Abstract*—Color is routinely present in health applications, but evidence about its role in older adults' user experience remains dispersed across design, health informatics, and human-computer interaction. This five-database study maps an affect-or-color focal corpus and applies a structured design-evidence audit to its color-term candidates. Searches of Web of Science Core Collection, Scopus, PubMed, IEEE Xplore, and the ACM Digital Library produced 2,515 source rows, 1,760 unique documents, and 941 focal records. A record-level audit of the classifier defining that corpus returned precision 0.880 and recall 0.982. Performance indicators, knowledge networks, and temporal bursts describe the corpus, and their stability was tested against retrieval, threshold, and periodization choices. Two design experts independently coded 41 color-term candidates, using full text for 23 records. Agreement reached 95.1% for inclusion (Cohen's kappa = 0.918) and 90.2% for population relevance (kappa = 0.853). Reconciliation retained ten primary older-adult studies, three mixed-age studies, and one contextual review; 27 records were excluded, and five of the ten color-and-emotion bridge records met the primary rule. Across the primary set, color supported legibility, risk communication, hierarchy, navigation, aesthetic preference, and affective tone, but most studies bundled color with other interface features. Preference-to-performance evidence was absent, and safety-relevant findings stayed embedded in broader workflows. The evidence map separates reported design functions from causal claims and identifies priorities for controlled validation.**

***Keywords*—Older adults; health applications; user experience; interaction design; color; emotion; bibliometrics; science mapping**

I. Introduction

Health applications mediate monitoring, consultation, education, and self-management for many older adults. Value rests on more than clinical content. Changes in vision, cognition, motor control, health literacy, and technology experience affect how information is found and acted upon. Wildenbos et al. locate these barriers within a mobile-health usability framework [1]. Acceptance studies add usefulness, compatibility, trust, and prior experience [2], [3]. Age-inclusive interaction design is therefore a task and evidence problem, not a matter of enlarging controls.

Color complicates that problem. It may guide attention, mark a status, organize a display, or alter the felt character of an interface. Its effect is conditional. Lightness and saturation relate to pleasure and arousal in controlled studies [7], [8].

Ecological and cross-cultural research emphasizes learned association and object context [9], [10], [11]. A palette that is liked cannot be assumed to improve comprehension, safe action, or continued use in a health task.

The relevant scholarship is distributed across health informatics, human-computer interaction, gerontology, medicine, service research, and design. Existing reviews usually remain inside one domain. They seldom connect publication growth and collaboration with the smaller body of studies that actually report color as an interface variable. A narrow conjunctive search would create the opposite problem by discarding the user-experience literature in which color questions are embedded. The gap is thus architectural as well as substantive.

This study addresses that gap with a two-layer corpus in which each layer keeps its own denominator. The outer layer is a focal map of 941 records. It describes production, collaboration, vocabulary, intellectual sources, and temporal fronts, and its task is to locate color work inside the surrounding user-experience literature rather than to supply design evidence. The inner layer is a structured design-evidence matrix. It examines 41 color-term candidates for age relevance, interface function, health task, perceptual or affective mechanism, behavioral outcome, and safety relevance, and it resolves to the 14 records that carry the study-level claims. The contribution lies in joining the two layers while holding their denominators apart, so that corpus-level visibility is never read as participant-level support. The analysis stops short of estimating clinical effectiveness and of proposing a universal palette. Four research questions organize the study.

RQ1: How have publication volume, source concentration, authorship, institutional participation, and country output developed in the retrieved field?

RQ2: Which collaboration, keyword, co-citation, and bibliographic-coupling structures organize its social, conceptual, and intellectual base?

RQ3: Which themes and citation fronts emerged or persisted across the study period?

RQ4: Within the color-term candidate set, which interface functions, health tasks, affective constructs, behavioral outcomes, attribution limits, and safety concerns are documented?

*Corresponding author.

II. Related Work and Analytical Position

A. Older-Adult User Experience in Health Applications

Older adults are not a uniform user group: functional ability, health literacy, device experience, and motivation all vary within the population. Studies of eHealth adoption therefore place usefulness, support, and trust alongside usability [2], [3], and digital literacy further conditions trust in online health information [19]. Older patients describe late-life learning demands as a barrier to eHealth use [20]. Interaction quality is relational: a visual feature acquires meaning through the task, the setting, and the person who encounters it.

Measurement remains inconsistent. Engagement reviews report many custom indicators and limited comparability across health applications [5]. Validated instruments such as uMARS, MAUQ, and the System Usability Scale provide more stable anchors [21], [22], [23], yet they leave color mechanisms unresolved, and a favorable usability score can coexist with inaccessible contrast or ambiguous alert semantics. The design question is more specific: which visual and interaction properties support a defined task for a defined user group?

B. Color, Emotion, and Affective Interaction Design

Color may distinguish controls, group information, signal status, reinforce hierarchy, support personalization, or establish an affective tone. These functions need separate treatment. Color-emotion relationships vary with lightness, saturation, object meaning, viewing conditions, and prior association [7], [8], [9], [10], [11]. Health interfaces add a semantic safety constraint. Warning, risk, and medication status should not depend on hue alone. The bibliometric question is therefore whether color is reported precisely enough to support cumulative user-experience knowledge.

C. Why a Bibliometric Map Is Necessary

Bibliometric analysis complements study-level review by revealing production, collaboration, conceptual vocabulary, and citation structure: co-word analysis locates recurring design problems, co-citation mapping identifies the intellectual base, and bibliographic coupling shows which current documents draw on similar references. Three evidence terms are kept distinct throughout. A metadata gap means a field is absent from the searchable export. A reporting gap means full-text verification finds the item missing from the study. An evidence gap means the available outcome evidence falls short of a design conclusion. A dense network testifies to citation and indexing behavior alone.

III. Data and Methods

A. Retrieval Design and Source Boundaries

Searches were executed on 31 July 2026 in Web of Science Core Collection, Scopus, PubMed, IEEE Xplore, and the ACM Digital Library. A single four-block Boolean expression operationalized the conceptual domains of older adults, health applications, user experience and interaction design, and color or affective terms. Every provider received the same four blocks, translated into its own field syntax. What differed between providers was field scope rather than concept coverage: Web of Science matched on Topic, Scopus on title, abstract, and keywords, IEEE Xplore on all metadata, PubMed on title and abstract, and ACM on abstract alone. Truncation appeared only in the fourth block, where emotion* and affect* were expanded by each provider's own stemmer. No date, language, or document-type limit was applied, so the retrieved set spans 2002 to 2026 and includes articles, reviews, and conference papers. Table I records the field scope, limit settings, and export boundary for each database, and Supplementary Material S2 reproduces the five expressions verbatim together with their result counts. A sensitivity analysis repeated the annual, source, keyword, and affiliation-country summaries after the nine PubMed-only records were removed. The study is a bibliometric corpus analysis with PRISMA-informed provenance rather than a systematic review of intervention effects [12].

TABLE I. Database Retrieval and Acquisition Boundaries

Database	Raw rows	Field scope, limits, and capture boundary
Web of Science Core Collection	1,308	Topic field; four blocks; stemmer truncation in block 4; no date, language, or type limit; executed 31 Jul 2026; full records with cited references
Scopus	687	Title, abstract, keywords; four blocks; stemmer truncation in block 4; no limits; executed 31 Jul 2026; result-list metadata
PubMed	356	Title/Abstract tags on all four blocks; stemmer truncation in block 4; no limits; executed 31 Jul 2026; XML and JSON
IEEE Xplore	118	All metadata; four blocks; stemmer truncation in block 4; no limits; executed 31 Jul 2026; result-list metadata
ACM Digital Library	46	Abstract field; four blocks; stemmer truncation in block 4; no limits; executed 31 Jul 2026; result-list metadata

Table I was reconstructed from the five saved result sets and their acquisition logs. Query scope appears beside export scope because those boundaries determine which fields can be pooled. All five providers ran the same four conceptual blocks; PubMed applied them within title and abstract, which narrows field scope without narrowing the concept set. Export depth differed more than query scope. Web of Science supplied full records and cited references, and PubMed supplied structured XML and JSON. Scopus, IEEE Xplore, and ACM records came largely from provider result lists, so abstracts, affiliations, keywords, and cited references were unevenly available. Table I therefore defines an analytical boundary rather than a list of interchangeable database samples.

Fig. 1 traces the provenance chain from the five database result sets to the study-level evidence tiers, following 2,515 source rows through 1,760 unique documents, 941 focal documents, and 41 color-term candidates to the resolved evidence tiers. Each stage carries a distinct denominator, which keeps metadata term occurrence from being mistaken for verified participant evidence.

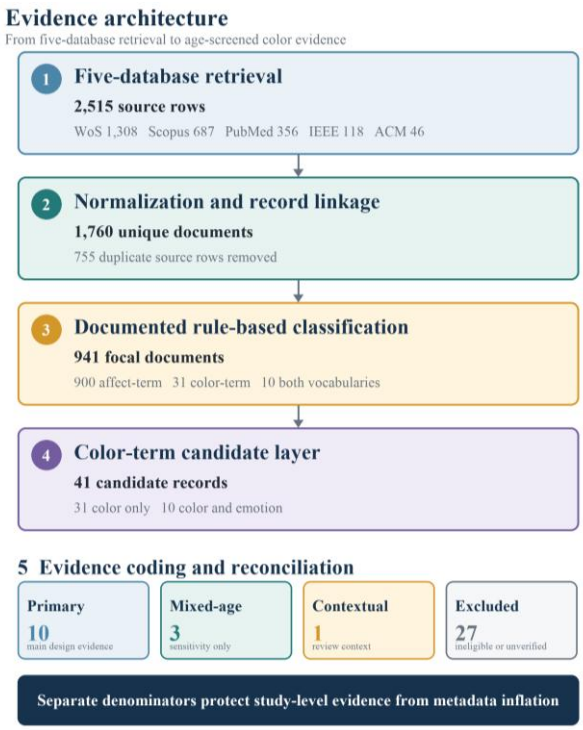

Fig. 1. Evidence architecture.

B. Normalization, Record Linkage, and Corpus Definition

OpenRefine 3.10.1 recorded 21 operations for whitespace, Unicode, identifiers, source labels, and document fields. DOI strings were lower-cased, stripped of resolver prefixes, and checked against title, year, and author evidence. Exact normalized titles were linked only when publication years differed by no more than one year. Unresolved candidates were checked against Crossref. A candidate was accepted only when title similarity reached 0.92, the year was compatible, and first-author evidence did not contradict it. The result contained 1,760 unique documents after 755 duplicate source rows were removed. Database coverage was one source for 1,315 documents, two sources for 140, and three sources for 305.

A documented rule-based classifier read titles, abstracts, author keywords, and index keywords. Color terms covered color, colour, chromatic, hue, lightness, brightness, palette, and color-qualified contrast or saturation. The affective vocabulary included emotion, affect, valence, arousal, trust, comfort, anxiety, enjoyment, frustration, fear, calmness, and satisfaction. The classifier found 900 records containing affect-related terms, 31 containing color-related terms, and 10 containing both vocabularies within the 941-record focal corpus. These labels describe metadata occurrence, not substantive research focus or complete article content. Age relevance was never inferred from a color term, an application name, or a demographic phrase such as people of color.

The classifier was then audited against the locked vocabulary across all 1,760 records rather than a sample, confirming the sense of each matched term and screening excluded records for terms the pattern could not reach. It returned 828 true positives, 113 false positives, 15 false negatives, and 804 true negatives: precision 0.880, recall 0.982, F1 0.928, and Cohen's kappa 0.855. Both error classes are systematic. Every false positive came from the verb affect, as in factors that affect adherence, which a lexical pattern cannot separate from the psychological construct; every false negative came from a morphological variant outside the word-boundary pattern.

Two consequences follow. The color stream is unaffected, since all 41 color-term candidates were confirmed and both error classes fall inside the surrounding affect vocabulary. The descriptive map also holds when the false positives are removed, retaining the ten leading sources and nineteen of the twenty leading keywords at rank correlations of 0.981 and 0.988, with the annual series at 0.995. The focal corpus is therefore an inclusive frame whose affect margin is documented rather than hidden (S5).

A further check addressed whether predefined terminology misses studies that describe color indirectly. The retrieval blocks already admit contrast, brightness, saturation, palette, trust, and engagement, so indirect descriptions enter at the search stage; the classifier vocabulary is narrower, requiring a color-qualified reading of contrast and saturation. An expanded probe adding unqualified contrast, visualization, display, theme, aesthetic preference, legibility, luminance, color-vision terms, named hues, and visual highlighting surfaced 546 further records, of which 111 survived sense disambiguation. None reported color as a manipulated or evaluated interface variable; the descriptors resolved to thematic-analysis themes, app-rating aesthetics subscales, data-dashboard visualization, and rhetorical uses. Seven records were kept for full-text verification (S2).

C. Bibliometric and Network Analyses

Bibliometrix 5.4.1 and Biblioshiny supported annual production, source and author-name-string productivity, thematic mapping, and thematic evolution [13]. VOSviewer 1.6.21 generated author, institution, and country collaboration maps, keyword co-occurrence, reference co-citation, and document-coupling networks [14], [15]. Fractional counting informed interpretation where large teams could otherwise dominate links [16]. Affiliation countries were parsed from all available address strings. Institution labels were case-folded and reconciled with a documented alias table; no ORCID-based or registry-based entity resolution was available.

TABLE II. NETWORK OBJECTS AND RETAINED ANALYTICAL MATRICES

Object	Nodes	Undirected edges
Author collaboration	250	663
Institution collaboration	250	547
Country collaboration	72	346
Keyword co-occurrence	250	11,494
Reference co-citation	300	6,427
Document coupling	941; 838 with cited references	20,161

Table II was generated from the retained adjacency matrices rather than copied from network images, and identifies the six network objects with their node and edge denominators. All six used full counting, association-strength normalization, and VOS clustering at resolution 1.00. Edge weight is the raw co-

occurrence, co-citation, or shared-reference count before normalization; nodes entered by rank on total link strength with occurrence weight as tie-breaker; isolates were dropped after truncation; and display subsets were pruned separately for single-column readability. Node size represents production or occurrence, line width link weight, and color the VOS cluster. Address data were available for 861 documents, affiliation-country fields for 855, institution fields for 839, and cited references for 838, so although the coupling matrix preserves all 941 rows, only those 838 records contribute nonzero coupling. S3 lists every parameter and retained node count by network.

Because structure responds to these choices, the six maps were re-clustered under varied parameters and compared with the published solution by adjusted Rand index, moving resolution between 0.7 and 1.3, substituting raw for association-strength weighting, and dropping the weakest quartile of nodes. Central nodes held throughout: the ten highest-strength nodes were retained in every configuration of all six networks. Cluster boundaries proved more sensitive and tracked density, holding at 0.80 to 1.00 for the sparse author and institution maps but ranging from 0.28 to 0.70 for the denser keyword, co-citation, coupling, and country maps, with community counts varying between two and eight. Semantic cluster membership is therefore reported as one defensible partition rather than a unique solution, with the interpretive weight on central nodes (S4).

D. Temporal Fronts and the Color-Explicit Design Matrix

Thematic evolution used three descriptive publication phases. Counted over complete years only, these hold 60 documents for 2009 to 2017, 267 for 2018 to 2022, and 437 for 2023 to 2025. The 177 records retrieved from January to July 2026 are reported separately as a partial observation and are excluded from every trend statement. The cut points mark the first year above 20 annual records and the later rise above 80, rather than equal-duration or equal-volume windows.

Because publication volume is a weak basis for claims about stages of development, both boundaries were tested against change points in the 2009 to 2025 series. A Pettitt test located a single change point at 2017 ($K = 72$, $p = 0.005$), adjacent to the first boundary. Two-breakpoint segmentation recovered 2023 on the raw-count scale, where the 2018 and 2023 pair ranked third among admissible pairs, but on a logarithmic scale the first boundary moved to 2012 or 2014, because the sparse years from 2009 to 2013 dominate that transformation. The second boundary is thus reproducible and the first scale-sensitive, so the three phases serve as descriptive reporting windows whose thematic labels summarize each window rather than mark a detected transition.

Keyword and cited-reference bursts were detected with Kleinberg's finite-state model [17]. The scale parameter was 2, the transition penalty was 1, the minimum keyword document count was 8, and the minimum cited-reference count was 5. DOI variants were merged before modeling. The procedure retained 198 keyword candidates and 286 cited-reference candidates. Because 2026 was incomplete, burst detection ended in 2025; the temporal corpus contained 764 documents, including 685 with cited references.

Two design experts independently coded all 41 color-term candidates using separate worksheets and a locked codebook. Fields covered population relevance, inclusion status, interface element, health task, color operation, proposed mechanism, affective construct, outcome, attribution, safety relevance, evidence location, and access status. Primary evidence required an explicit older-adult designation supported by participant age information or an older-adult user test, a health-related digital interface, and explicit color evidence. A record was mixed-age when older participants or a documented older subgroup were present without a separate older-adult result. Target-only expert studies and reviews were classified as contextual rather than primary. Inclusion decisions agreed for 39 of 41 records (95.1%; Cohen's kappa = 0.918; bootstrap 95% CI 0.80 to 1.00). Population relevance agreed for 37 of 41 (90.2%; kappa = 0.853; 95% CI 0.70 to 0.97). A third reviewer reconciled six records containing a disagreement in either field against the most complete accessible source. Secondary categorical kappa values ranged from 0.705 to 1.000; overlapping open-text design fields remained descriptive.

E. Quality Controls and Reproducibility

Quality control was applied at the source, record, corpus, and result levels. Source counts were reconciled against the five saved result sets before parsing, and each raw row retained its database name, provider identifier, source file, and URL. Field-completeness audits were run separately for DOI, title, author, abstract, keyword, affiliation, citation, and cited-reference data, so missingness was visible before any network was built. Optional fields had no bearing on corpus membership. Analytical outputs were regenerated from the retained corpus rather than transcribed from images, and each figure is linked to its map, network, or tabular source file.

The linkage audit recorded every DOI union and exact-title union, together with the Crossref score and DOI origin. A resolver check identified the provider DOI 10.2024/1/e52442 in cluster CE-00953 as non-resolving. Title, author, year, and Crossref metadata confirmed the canonical DOI 10.2196/52442, which is the identifier used for interpretation in the corrected audit. This cluster contains four source rows from Web of Science, Scopus, and PubMed but represents one document. It is not a legitimate reprint and is not counted twice.

One interpretive boundary was fixed before the maps were read. Node prominence measures bibliometric visibility, and its warrant stops there; design and clinical effectiveness lie outside what it can register. Together with the gap vocabulary of Section II-C, this keeps the quantitative map within the limits of its evidence.

IV. Results

The results follow the two layers of the corpus. Sections IV-A to IV-D describe the 941-record focal map and the user-experience literature surrounding color work, each stating what its finding implies for the color layer; Section IV-E then resolves that layer from the 41 color-term candidates to the 14 records carrying study-level evidence. Denominators differ at every step and are named wherever a count appears.

A. Corpus Formation and Growth

The 2,515 source rows contracted to 1,760 unique documents, a retention rate of 69.98%. Within the corpus, 1,311 documents came from one raw source row, 144 from two, 304 from three, and one from four; this differs from the database-coverage distribution in Section III-B because some documents were duplicated inside a single database. The CE-00953 DOI was canonicalized to 10.2196/52442 after resolver and Crossref checks, and no document was retained as a separate reprint on the basis of the invalid provider DOI.

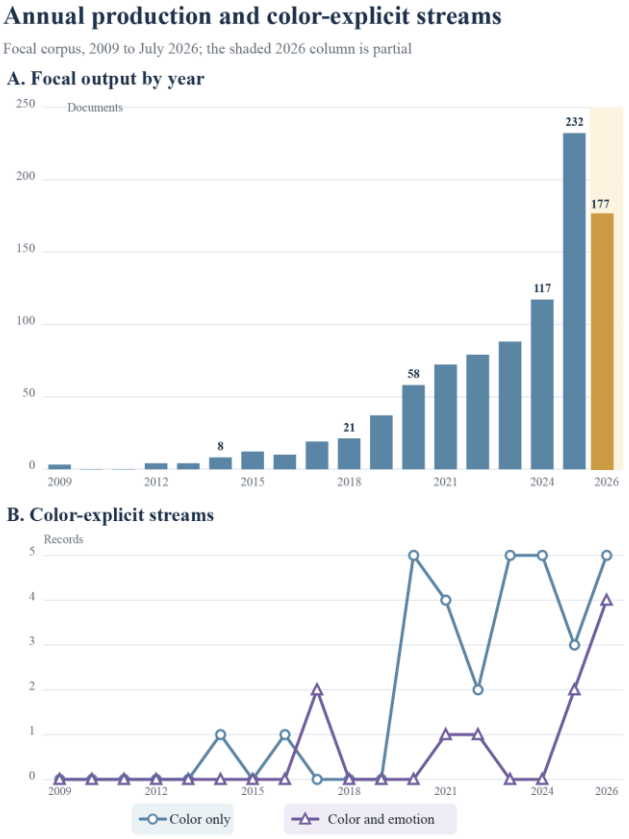

Fig. 2. Annual production and color-term streams.

Fig. 2 comes from the deduplicated annual-production table. Panel A shows the scale of the focal corpus; Panel B retains the smaller color-term streams on their own axis. Annual focal output rose from 21 documents in 2018 to 58 in 2020, 117 in 2024, and 232 in 2025. The 177 records retrieved from January to July 2026 cover seven months and are shaded as a partial observation. Every month, they run at 25.3 records against 19.3 for 2025, which annualizes to roughly 303. That comparison shows only that the partial year is not evidence of decline; because submission and indexing are uneven within a year, the 2026 figure stays out of every growth statement and out of burst detection. Excluding nine PubMed-only records left 932 documents and changed any annual count by no more than three. The top ten source set and top twenty keyword set were unchanged; frequency rank correlations were 0.980 and 0.994, respectively. Nine of the ten leading affiliation countries remained, with rho = 0.996. The sensitivity result supports the stability of the descriptive map without making the PubMed query prevalence-equivalent.

B. Sources, Contributors, and Collaboration

Case-folding and alias inspection reduced 407 nonblank source labels to 399 normalized labels. JMIR Formative Research contributed 101 focal documents. The Journal of Medical Internet Research contributed 63, JMIR Research Protocols 41, JMIR mHealth and uHealth 36, JMIR Human Factors 35, and both Digital Health and JMIR Aging 26. The source core belongs mainly to digital health and health informatics. General HCI and design venues form a smaller visible layer, although the retrieval design cannot establish what lies beyond the corpus.

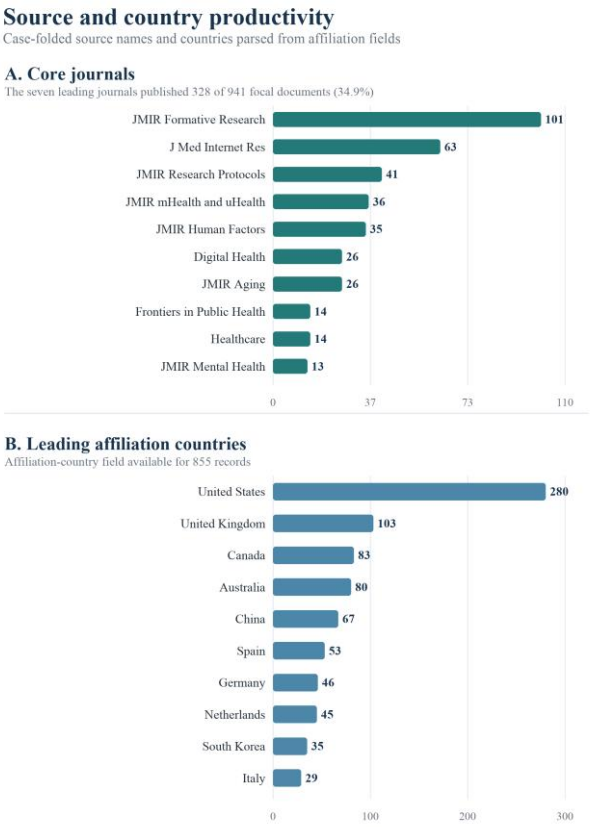

Fig. 3. Core journals and leading affiliation countries.

Fig. 3 places source concentration beside country participation parsed from affiliation addresses. The panels use different metadata fields. The United States appeared in the affiliations of 280 documents, followed by the United Kingdom in 103, Canada in 83, Australia in 80, and China in 67; these are affiliation-presence rather than corresponding-author counts. Institution strings were normalized separately. The University of Toronto led with 19 documents, the University of Sydney with 14, and the University of British Columbia with 13. These values describe recorded participation, whereas Fig. 4 examines collaboration strength.

After source-name normalization, the seven leading sources accounted for 328 documents, or 34.86% of the focal corpus, while 323 of the 399 normalized sources contributed one document. Authorship was more dispersed: 6,969 normalized author-name strings were recorded, of which 6,722 appeared once. Institution strings showed a similar long tail, with 2,014 of

2,262 occurring in one document. The pattern indicates dispersion at the recorded-name level.

Those institution counts rest on string identity, so the leading entities were audited against the address field, which exposed two opposite failures. Splitting is the larger: 1,076 of the 2,262 institution strings retain a fragment of the bracketed author list preceding each Web of Science address, so one institution can appear under several labels. Re-parsing and crediting each document once raises the leading counts to 26, 22, and 20 for Toronto, Sydney, and British Columbia without changing their order. Merging is the opposite risk, since a token-insensitive normalization would combine Univ Washington in Seattle with Washington Univ in St. Louis, which the city field separates. Across the published top thirty, the two parses correlate at rho = 0.759 and seven of ten identities persist, so the three leading institutions are stable while positions below them turn on identity decisions string matching cannot settle (S6).

This distribution bears directly on the color layer. A concentrated core coexists with a long tail of single-study contributors, and across that tail color surfaces under varied labels: contrast, theme, display, visualization, palette, or aesthetic preference. The vocabulary probe reported in Section III measured that dispersion, finding 111 records that use such descriptors while carrying no color term in their metadata. Section V returns to what the pattern implies for reporting practice.

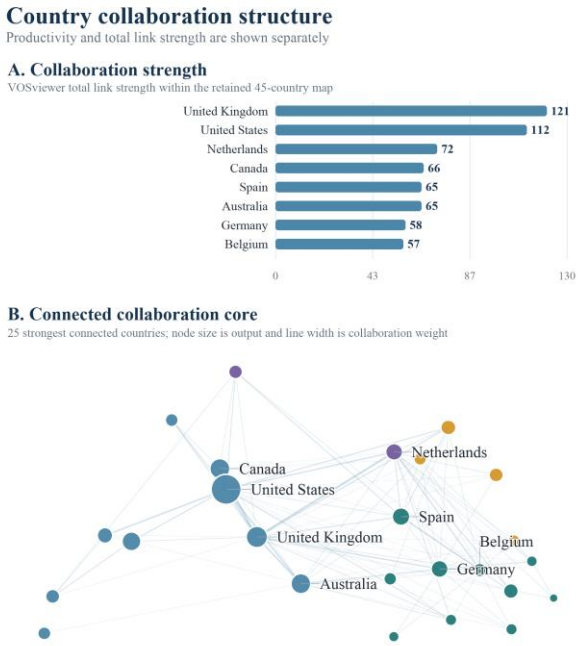

Fig. 4. Country productivity and collaboration structure.

Fig. 4 separates publication output from network connection. Panel A reports total link strength in the retained 45-country VOSviewer map; Panel B shows the 25 strongest connected countries and their 100 strongest links. The United Kingdom had the highest total link strength at 121, followed by the United States at 112. The Netherlands reached 72, Canada 66, Spain and Australia 65 each, Germany 58, and Belgium 57. The University of Toronto led the institutional network with total link strength 41, while Bucci and Schwannauer each reached 64 in the author network. These values refer only to the retained maps and available affiliation fields. Having distinguished productivity from connection, the analysis turns to the vocabulary that structures the field.

C. Conceptual Structure and Thematic Evolution

Raw author-keyword frequencies were led by mHealth with 199 documents, digital health with 174, mobile health with 118, usability with 111, eHealth with 109, and older adults with 95. Mental health, depression, anxiety, telehealth, acceptance, user experience, and feasibility formed the next layer. Color terms remained outside the dominant raw vocabulary. Because mHealth and mobile health, as well as eHealth and digital health, are related but not identical labels, the analysis reports them separately and interprets their combined design meaning cautiously.

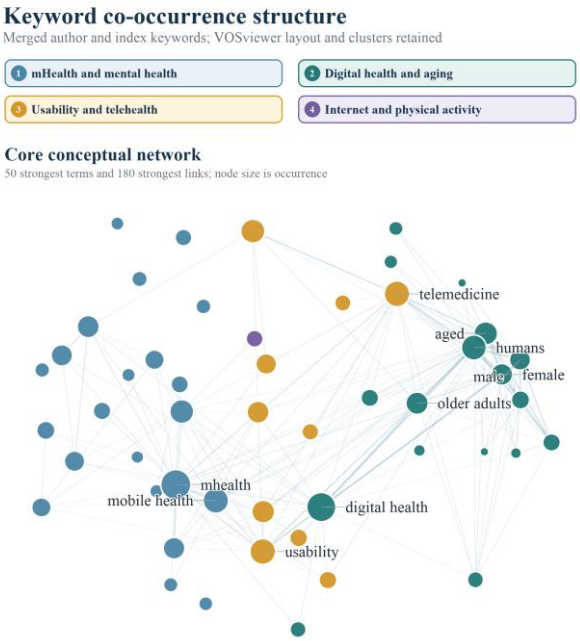

Fig. 5. Keyword co-occurrence structure with named conceptual clusters.

Fig. 5 was reconstructed from the saved VOSviewer map using the 50 terms with the greatest total link strength and the 180 strongest links. Four communities organize the visible vocabulary. The first connects mHealth and mobile health with mental health, depression, anxiety, and intervention. The second links digital health with older adults and demographic indexing. A third community brings usability, telemedicine, self-management, and acceptance together. The smaller fourth group concerns internet use and physical activity. Cluster names above the network state the interpretation in words, reducing reliance on color alone. Color terminology again stays outside the dominant keyword core.

The thematic map refines this reading without assigning causal status to clusters. The eHealth, depression, and anxiety cluster has the highest relative centrality and density; mHealth and usability are central but less internally specialized; digital health and aging show developing connectivity; and telehealth

care occupies a less central, less dense position. These positions describe how terms are organized in the retrieved literature, leaving effectiveness to study-level evidence.

Thematic evolution shows a change in the vocabulary used by the retrieved literature. From 2009 to 2017, labels centered on technology, telemedicine, mobile health, telehealth, and intervention. During 2018 to 2022, the vocabulary consolidated around mHealth, telemedicine, and anxiety. From 2023 onward, digital health and usability became more explicit destination themes. Word-flow links early telemedicine to later digital-health and implementation vocabulary. The pattern marks greater visibility of interaction-quality terms, not evidence of theoretical maturation.

D. Intellectual Base and Current Coupling Fronts

The co-citation base combines usability measurement, technology acceptance, clinical screening, research infrastructure, and digital-mental-health guidance. Representative references include the System Usability Scale [23], the Unified Theory of Acceptance and Use of Technology [24], GAD-7 [25], REDCap [26], and work on engagement with mental-health applications [27]. These references form the methodological backbone of the map. They show that the field draws on HCI, information systems, psychometrics, and health informatics rather than on one established color-design tradition.

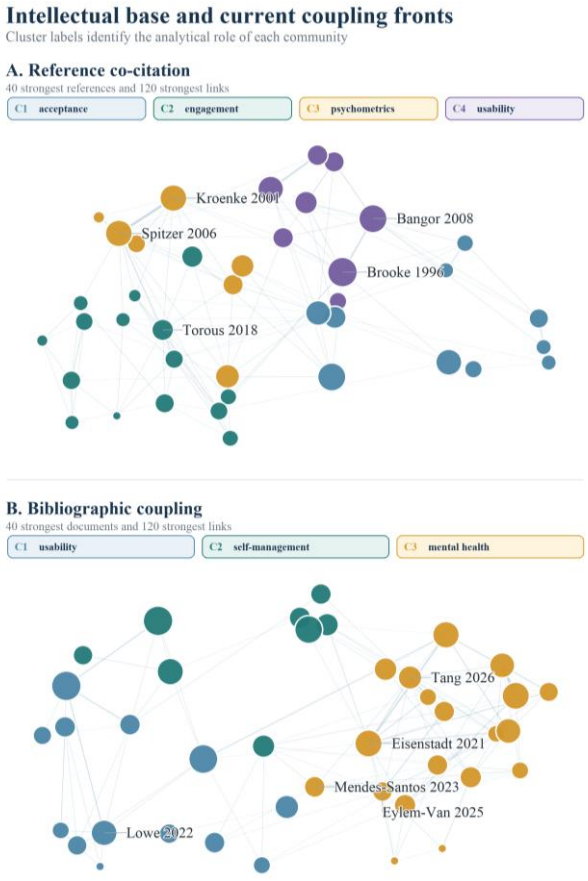

Fig. 6. Co-citation and document-coupling structures.

Fig. 6 distinguishes the intellectual base from current coupling fronts. Panel A contains four co-citation communities centered on acceptance, engagement, psychometrics, and usability. Representative references include UTAUT, GAD-7, engagement guidance, and the System Usability Scale [23], [24], [25], [27]. Panel B contains three displayed document-coupling communities concerned with usability, self-management, and mental health. Short labels identify the documents actually retained in the single-column map. Telehealth satisfaction and application-based physical-activity research belong to the wider coupling corpus [28], [29], but are not among its displayed labels. The structure is grounded in measurement and adoption rather than a mature color-design tradition.

Citation counts provide a complementary snapshot, taken as the maximum provider count in the saved July 2026 exports. Donker et al.'s systematic review of mental-health applications led with 774 citations [4]. Li et al.'s older-adult wearable acceptance model followed with 328, Lehtimaki et al.'s digital-mental-health overview with 302, and Ng et al.'s engagement review with 200 [2], [5], [6]. Fan et al.'s real-world chatbot study also ranked highly [18]. Being neither field-normalized nor cross-provider equivalent, these values support ranking only within the assembled corpus.

Co-citation and coupling answer different questions. The co-citation map reflects references that later papers repeatedly place together, so its clusters indicate a shared intellectual base. Coupling depends on the reference lists of current documents and responds more quickly to new problems. The recent coupling front contains a visible 2024 to 2026 layer even though those records have had less time to accumulate citations. That layer sits apart from the historical base in the map, a separation of reference profiles rather than a measure of influence.

E. Temporal Bursts and the Color-Explicit Design Subset

Sixteen keyword burst intervals and seven DOI-normalized cited-reference bursts were detected. Early keyword bursts were broad and persistent: intervention ran from 2009 to 2020 with strength 6.23, internet from 2009 to 2017 with strength 4.47, and technology from 2012 to 2016 with strength 4.22. Later bursts became more methodological. Validation ran from 2020 to 2023 with strength 5.83, while mobile phone ran from 2021 to 2023 with strength 5.25. Female, male, artificial intelligence, and qualitative research appeared as short bursts in 2025. These are indexing and publication fronts, not direct evidence of design effectiveness.

The normalized reference bursts trace a parallel methodological shift. A mobile application rating reference peaked in 2022 [21]. GAD-7 burst from 2020 to 2022, and the digital mental-health engagement review from 2021 to 2022 [25], [27]. REDCap rose in 2023 and 2024, showing increased citation visibility for standardized data infrastructure [26]. UTAUT produced a 2024 burst, marking renewed attention to adoption theory [24]. Two shorter bursts concerned application assessment and public-health implementation. The System Usability Scale did not remain a burst after DOI variants were collapsed, correcting the earlier duplicate-sensitive result.

The burst sequence moves through three emphases. Early fronts concerned whether mobile technology could deliver an intervention. Middle-period fronts emphasized acceptability, psychometrics, and validation. The latest fronts concern how

digital services are implemented and experienced by more specific user groups. Color formed no independent burst at any stage, and its terms stayed embedded in usability, visualization, accessibility, and experience studies.

Fig. 7. Temporal fronts and age-screened color evidence.

Fig. 7 brings temporal fronts and design evidence into one view while keeping their denominators separate. Panel A plots the ten strongest keyword bursts from the complete-year corpus through 2025. Panel B shows the evidence tiers derived from 41 color-term candidates, followed by five nonexclusive characteristics of the ten primary records: full-text availability, color isolation, direct safety relevance, preference-to-performance linkage, and isolated safety endpoints. The panel is an evidence-boundary summary rather than a causal-effect display.

TABLE III. AGE RELEVANCE AND EVIDENCE TIERS IN THE COLOR TERM CANDIDATE SET

Boundary	Result and interpretation
Primary older-adult	10 (6 full text; 4 abstract): main design evidence
Mixed-age evidence	3 (2 full text; 1 abstract): no separate older result
Contextual	1 abstract-level scoping review
Excluded	27 of 41: 24 noneligible, 2 contextual-only, 1 unverified
Primary bridge	5 of 10; no fully isolated color-only effect

Table III formalizes the boundary shown in Fig. 7. Ten records met the primary older-adult criterion, including six coded from full text and four from the most complete abstract-level record. Three mixed-age studies entered sensitivity interpretation; two were supported by full text and one by abstract-level evidence. One abstract-level scoping review remained contextual. The other 27 records comprised 24 noneligible studies, two target-only contextual exclusions, and one unverified title-level record. This tiering prevents the 41-record candidate layer from being interpreted as a homogeneous older-adult evidence set. Study-level fields for all 14 retained records are provided in Supplementary Table S1.

Five primary color-and-emotion bridge records addressed distinct design functions. Falls-risk feedback combined color with text, icons, and texture [31]. Multimorbidity telemonitoring used a multicolor risk indicator within a complete risk-and-notification workflow [32]. An ePHR study modeled color and icon features as a bundled predictor of usefulness, ease of use, and intention [33]. The A-Kano study embedded bright color within a broader emotional-design package [34]. In the eye-tracking experiment, navigation depth explained more than color, although blue supported more efficient visual exploration [35]. The remaining primary studies examined interface preference [36], accessible color information [37], medication-support themes [38], telepresence contrast [42], and patient-portal legibility for older migrants [44]. Mixed-age evidence concerned graph framing, tuberculosis support, and caregiver self-care [39], [40], [41]. One scoping review remained contextual [43].

Across the resolved matrix, legibility and redundant information were recurring design concerns [31], [37], [42], [44]. Other studies used color for risk status, information hierarchy, navigation, aesthetic preference, or affective tone [32], [34], [35], [36], [38]. These functions imply different mechanisms. Contrast may affect detection, categorical color may affect comprehension, and a palette may influence stated preference. Combining them under one generic color effect would obscure both task and risk.

The causal evidence remained thin. Eight of the ten primary records coded color as not isolated, one as partially isolated, and one as unclear. Nine did not provide an interpretable preference-to-performance test; the remaining study tested the link but did not support it. Five records were directly relevant to safety-critical use, although their outcomes belonged to broader interface or service workflows. None produced a fully isolated color-safety estimate.

V. DISCUSSION

A. A Broad User-Experience Field with a Thin Color Layer

The focal corpus reveals a recognizable bibliometric structure: an international literature concentrated in a small source core and organized around usability, mental health, telehealth, acceptance, and digital service delivery. Its color-specific layer is much smaller than the surrounding affect vocabulary. Resolving candidate records at study level therefore changes the interpretation from topic visibility to documented design evidence. That distinction prevents a prominent metadata term from being treated as participant-level support.

The temporal record helps explain that thinness. Color never formed an independent keyword or citation front; its terms stayed embedded in usability, visualization, accessibility, and experience studies. A variable reported only inside something else accumulates neither shared terminology nor comparable measurement, which is the condition the evidence matrix meets at study level.

The primary studies show why color is better understood through interface function than through a fixed emotional symbol. Accessibility work emphasizes contrast and redundancy [31], [37], [42], [44], while preference and acceptance studies describe stated utility or intention [33], [36], [38]. Risk-status work places color inside a larger decision workflow [32], and the eye-tracking evidence concerns search and trust, with navigation structure explaining more than color [35]. Affective constructs are often invoked without a validated affect measure, and color is rarely manipulated alone. The literature therefore supports testable design hypotheses and reporting standards rather than a claim that one hue is inherently calming, trustworthy, or safe for older users.

B. Implications for Design Evidence and Reporting

Future studies should define the function of color before testing an interface, because each function implies a different design and a different endpoint. A legibility manipulation requires luminance contrast, typography, viewing distance, and task-performance outcomes. A semantic-alert study requires redundant coding, error costs, and comprehension measures. An affective manipulation requires specified color coordinates and a validated emotional outcome.

TABLE IV. PROPOSED RESEARCH RECOMMENDATIONS

Observed reporting issue	Minimum design specification	Proposed validation outcome
Color not quantified	Report color space and coordinates	Replicable visual condition
Contrast absent	Report luminance ratio and text size	Recognition, error, completion time
Affect weakly measured	Use a validated affect measure	Valence, arousal, trust, or stress
Context omitted	State device, illumination, culture	Transferability across settings
Color carries meaning	Add text, shape, icon, or position	Safety and comprehension

Table IV translates the reconciled evidence matrix and its reporting gaps into proposed research recommendations, and reproducibility comes first among them. A named hue such as blue cannot define an experimental condition when displays render different luminance and saturation. A reproducible record specifies the color space and coordinates, foreground and background, contrast ratio, text size, device and screen settings, and ambient illumination; screenshots may supplement those values but cannot replace them. Because older adults meet the same interface under different screens, lighting, and visual capabilities, that context belongs in the record too. WCAG 2.2 offers an accessibility reference [30], yet conformance stops short of empirical testing with intended users, and cultural context deserves equal attention because color meaning is learned and situated.

Outcome alignment matters just as much. Legibility claims need recognition, error, completion-time, or visual-attention evidence, and affective claims need a validated measure of valence, arousal, trust, or anxiety. Adoption, preference, and satisfaction remain useful secondary outcomes, but they cannot stand in for perceptual performance or safety evidence. Critical meaning needs redundancy for the same reason: color can carry hierarchy and rapid scanning, yet a critical state should stay legible through text, shape, icon, position, texture, or sound whenever color perception is uncertain. A rigorous evaluation tests that redundant representation instead of merely mentioning it.

C. Methodological Implications for Bibliometric Studies

Corpus architecture shapes what the analysis can claim. An intersection search would hide the surrounding user-experience literature, whereas a single undifferentiated corpus would let broad affect vocabulary obscure the smaller color-specific layer. Separate denominators preserve context without inflating design evidence, and the same architecture can support other narrowly reported interface variables.

Provider heterogeneity nevertheless limits comparability. All five databases ran the same four conceptual blocks, but field scope and export depth varied, so citation and affiliation coverage differ by provider. Removing the nine PubMed-only records produced only small changes in annual counts and ranking structure. The map is stable to that difference while remaining an assembled corpus rather than a field-neutral union of databases.

D. Toward a Cumulative Evidence Program

A cumulative program begins with descriptive standardization. Authors can name the function of color, state whether it was manipulated or merely present, and report the values that reproduce it, adding such fields to existing checklists without requiring a new theory of color. Retrieval depends on the same discipline: color is often visible in an interface yet invisible in its bibliographic description, so abstracts and keywords should identify color wherever it is manipulated, evaluated, or used to carry meaning, with structured supplementary fields holding the numerical detail.

At a more demanding level, studies need mechanism isolation. When palette, contrast, iconography, wording, and layout change together, no outcome can be attributed to color. A stronger experiment varies one color property while holding semantic content and geometry stable, and factorial designs can test whether lightness, saturation, or hue interacts with task urgency, age-related vision, or affective framing. Manipulation checks should confirm that participants perceived the intended difference before preference or performance is interpreted.

The outcome layers should also be kept apart. Perceptual outcomes include detection, discrimination, reading speed, and visual search; behavioral outcomes include errors, completion, adherence, and recovery from mistakes; affective outcomes include valence, arousal, trust, anxiety, and comfort; clinical outcomes usually require longer follow-up. Separating them keeps a favorable preference rating from being read as evidence of safer or more effective use.

Age heterogeneity deserves its own analysis, since chronological age describes neither contrast sensitivity, color discrimination, and digital literacy nor health status and cultural interpretation. Studies should record the user characteristics theoretically linked to the visual task. Samples rarely support every subgroup, yet planned stratification and transparent uncertainty are preferable to a single assumed older-adult response, and cross-cultural replication matters most where color carries evaluative or risk meaning.

Laboratory evidence also needs an ecological bridge. Controlled screens provide precision rather than the variable illumination, distraction, fatigue, and device settings of daily use. Promising effects should move to field deployment with predefined stopping rules for safety-critical interfaces, and longitudinal work can show whether an initial affective advantage sustains use or fades with novelty.

Such a program also depends on collaboration. The country network shows capacity for international work, and source concentration identifies journals able to normalize better reporting, so design researchers can specify visual variables, clinicians define error costs, and human-factors specialists align outcomes with tasks. The present color matrix offers a starting codebook for that shared language.

For practitioners, the evidence supports a conservative principle that follows the same function-first logic. Endpoints should match the function color serves: search time, route errors, and recall for navigation and hierarchy; detection, comprehension, and error severity for alerts or medication status; and measures separating immediate liking from trust, comfort, or sustained engagement for affective uses. Color can strengthen grouping, salience, identity, and tone, but it should never be the sole carrier of critical information, so design systems should record accessible defaults, redundant cues, tested display conditions, and the circumstances justifying a visual variant.

The bibliometric maps indicate where such standards could diffuse. A small group of sources publishes a large share of the literature and international collaboration is substantial, which makes the source core a plausible site for reporting improvement, though concentration alone says nothing about design quality. The long tail matters equally, since terminology drift is strongest there: a study calling its manipulation a theme, a display setting, or an aesthetic preference leaves no durable keyword, and later retrieval loses the record.

VI. Limitations

The first limitation is unequal search and export depth. As Table I records, Web of Science supplied full records and cited references; Scopus, IEEE Xplore, and ACM were represented largely by result-list metadata; and PubMed matched within title and abstract. Country, institution, citation, and reference networks are consequently shaped more strongly by records with richer fields and by documents found in several databases. The PubMed-only sensitivity analysis showed stable leading sources and keywords, though the five interfaces remain distinct instruments, and the maps describe the assembled corpus rather than each provider's complete native network.

Record linkage introduces a separate qualification. DOI and conservative title evidence were used, and the corrected audit replaces the nonresolving provider DOI in cluster CE-00953 with 10.2196/52442. Records without stable identifiers may still remain separate, and citation values are maximum provider counts captured in July 2026 without field normalization. Author and institution analyses rest on normalized strings rather than ORCID or registry identifiers. The entity audit in Section IV.B shows the cost: address-field fragments split institutions, plausible normalizations can merge distinct ones, and the two parses correlate at rho = 0.759 across the published top thirty. The three leading institutions survive both parses, so rankings deserve confidence at the top and caution below it.

The color matrix has its own constraints. Full text was available for 23 of 41 candidates, so the remaining 17 abstract-level records and one title-level record support less detailed inference. Polysemy and incomplete indexing can alter discovery. The coding protocol improved reproducibility for population and inclusion decisions, yet agreement under one codebook establishes consistency rather than causal validity. The classifier audit adds a second boundary: precision of 0.880 marks the affect stream as deliberately inclusive, and the verb-sense false positives stay identified in S5, so their effect remains traceable. That margin leaves the color layer untouched.

Time and modeling choices add further uncertainty. The 2026 records cover January to July, are reported separately on a monthly basis, and stay outside burst detection. The three thematic periods have unequal duration and volume, because they represent publication phases rather than comparable exposure windows. Change-point testing reproduces the 2023 boundary but places the first anywhere between 2012 and 2018 depending on scale, so the phases function as reporting windows rather than detected regimes. Network appearance likewise depends on thresholds, counting, and layout; the stability analysis separates central nodes, which held in every configuration, from cluster boundaries in the denser semantic maps, which moved with resolution. Records with unverified age or interface-color evidence were excluded rather than imputed, trading sensitivity for interpretability. Future updates should obtain the remaining full texts, preregister the coding protocol, and test whether an external team reproduces the resolved classifications.

VII. Conclusion

This five-database study maps 941 focal records around older adults' health applications, user experience, interaction design, color, and emotion. Output rose sharply, a digital-health source core became visible, and collaboration extended across countries, while usability, mental health, telehealth, acceptance, and service implementation organized the conceptual and intellectual structures. Auditing showed where those descriptions hold: central network nodes and the leading institutions survived changes of parameter and parsing, whereas semantic cluster boundaries and the earlier phase boundary did not. The temporal results indicate growth and differentiation; theoretical and methodological maturity remains an open question.

Color forms a small and uneven evidence layer inside that corpus. The resolved matrix retained ten primary older-adult studies, three mixed-age studies, and one contextual review, and only five color-and-emotion bridge records qualified as primary. Reported functions ranged from legibility and risk communication to navigation, preference, and affective tone, yet most studies changed several interface features at once, so the links from preference to performance and from color to safety remain unresolved. Stronger evidence requires reproducible color values, a defined health task, redundant critical cues, measures matched to the proposed mechanism, and explicit limits on causal attribution.

SUPPLEMENTARY MATERIALS

All supplementary files are openly available through the online repository at https://doi.org/10.5281/zenodo.22798290. Supplementary Table S1 gives the study-level evidence matrix for the 14 retained records. Supplementary Material S2 reproduces the five database expressions with their field tags, limits, execution dates, and result counts, and reports the expanded-vocabulary recall probe. S3 lists the network construction parameters, S4 the stability matrix and change-point tests, S5 the record-level classifier audit, and S6 the entity-identity audit.

DATA AVAILABILITY

The same repository record also holds the search strategies, derived datasets, coding materials, reliability results, and figure source data, and these remain available from the first author on reasonable request. No raw provider export is redistributed, because database licences may restrict redistribution of abstracts and record-level fields.

DECLARATION ON GENERATIVE ARTIFICIAL INTELLIGENCE

The authors used generative artificial intelligence tools to assist with language refinement and with coding the data-processing and analysis scripts used in this study. All such scripts and their outputs are openly available in the repository record cited above. The authors verified every reported result and take full responsibility for the content of this article.

REFERENCES

[1] G. A. Wildenbos, L. Peute, and M. Jaspers, "Aging barriers influencing mobile health usability for older adults: A literature-based framework (MOLD-US)," International Journal of Medical Informatics, vol. 114, pp. 66-75, 2018, doi: 10.1016/j.ijmedinf.2018.03.012.

[2] J. Li, Q. Ma, A. H. Chan, and S. S. Man, "Health monitoring through wearable technologies for older adults: Smart wearables acceptance model," Applied Ergonomics, vol. 75, pp. 162-169, 2019, doi: 10.1016/j.apergo.2018.10.006.

[3] A. Ghorayeb, R. Comber, and R. Gooberman-Hill, "Older adults' perspectives of smart home technology: Are we developing the technology that older people want?," International Journal of Human-Computer Studies, vol. 147, p. 102571, 2021, doi: 10.1016/j.ijhcs.2020.102571.

[4] T. Donker et al., "Smartphones for Smarter Delivery of Mental Health Programs: A Systematic Review," J Med Internet Res, vol. 15, no. 11, p. e247, 2013, doi: 10.2196/jmir.2791.

[5] M. M. Ng, J. Firth, M. Minen, and J. Torous, "User Engagement in Mental Health Apps: A Review of Measurement, Reporting, and Validity," PS, vol. 70, no. 7, pp. 538-544, 2019, doi: 10.1176/appi.ps.201800519.

[6] S. Lehtimaki, J. Martic, B. Wahl, K. T. Foster, and N. Schwalbe, "Evidence on Digital Mental Health Interventions for Adolescents and Young People: Systematic Overview," JMIR Ment Health, vol. 8, no. 4, p. e25847, 2021, doi: 10.2196/25847.

[7] A. J. Elliot and M. A. Maier, "Color Psychology: Effects of Perceiving Color on Psychological Functioning in Humans," Annual Review of Psychology, vol. 65, no. 1, pp. 95-120, 2014, doi: 10.1146/annurev-psych-010213-115035.

[8] P. Valdez, and A. Mehrabian, "Effects of color on emotions," Journal of Experimental Psychology: General, vol. 123, no. 4, pp. 394-409, 1994, doi: 10.1037/0096-3445.123.4.394.

[9] S. E. Palmer, and K. B. Schloss, "An ecological valence theory of human color preference," Proc. Natl. Acad. Sci. U.S.A., vol. 107, no. 19, pp. 8877-8882, 2010, doi: 10.1073/pnas.0906172107.

[10] L. C. Ou, M. R. Luo, A. Woodcock, and A. Wright, "A study of colour emotion and colour preference. Part I: Colour emotions for single colours," Color Research & Application, vol. 29, no. 3, pp. 232-240, 2004, doi: 10.1002/col.20010.

[11] L. C. Ou, M. R. Luo, A. Woodcock, and A. Wright, "A study of colour emotion and colour preference. Part II: Colour emotions for two-colour combinations," Color Research & Application, vol. 29, no. 4, pp. 292-298, 2004, doi: 10.1002/col.20024.

[12] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," PLoS Med, vol. 18, no. 3, p. e1003583, 2021, doi: 10.1371/journal.pmed.1003583.

[13] M. Aria, and C. Cuccurullo, "bibliometrix : An R-tool for comprehensive science mapping analysis," Journal of Informetrics, vol. 11, no. 4, pp. 959-975, 2017, doi: 10.1016/j.joi.2017.08.007.

[14] N. J. v. Eck, and L. Waltman, "Software survey: VOSviewer, a computer program for bibliometric mapping," Scientometrics, vol. 84, no. 2, pp. 523-538, 2010, doi: 10.1007/s11192-009-0146-3.

[15] N. J. v. Eck, L. Waltman, R. Dekker, and J. v. d. Berg, "A comparison of two techniques for bibliometric mapping: Multidimensional scaling and VOS," J. Am. Soc. Inf. Sci., vol. 61, no. 12, pp. 2405-2416, 2010, doi: 10.1002/asi.21421.

[16] A. Perianes-Rodriguez, L. Waltman, and N. J. v. Eck, "Constructing bibliometric networks: A comparison between full and fractional counting," Journal of Informetrics, vol. 10, no. 4, pp. 1178-1195, 2016, doi: 10.1016/j.joi.2016.10.006.

[17] J. Kleinberg, "Bursty and hierarchical structure in streams," Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 91-101, 2002, doi: 10.1145/775047.775061.

[18] X. Fan et al., "Utilization of Self-Diagnosis Health Chatbots in Real-World Settings: Case Study," J Med Internet Res, vol. 23, no. 1, p. e19928, 2021, doi: 10.2196/19928.

[19] S. R. Paige, J. L. Krieger, and M. L. Stellefson, "The Influence of eHealth Literacy on Perceived Trust in Online Health Communication Channels and Sources," Journal of Health Communication, vol. 22, no. 1, pp. 53-65, 2017, doi: 10.1080/10810730.2016.1250846.

[20] V. M. Nymberg et al., "'Having to learn this so late in our lives…' Swedish elderly patients' beliefs, experiences, attitudes and expectations of e-health in primary health care," Scandinavian Journal of Primary Health Care, vol. 37, no. 1, pp. 41-52, 2019, doi: 10.1080/02813432.2019.1570612.

[21] S. R. Stoyanov, L. Hides, D. J. Kavanagh, and H. Wilson, "Development and Validation of the User Version of the Mobile Application Rating Scale (uMARS)," JMIR Mhealth Uhealth, vol. 4, no. 2, p. e72, 2016, doi: 10.2196/mhealth.5849.

[22] L. Zhou, J. Bao, I. M. A. Setiawan, A. Saptono, and B. Parmanto, "The mHealth App Usability Questionnaire (MAUQ): Development and Validation Study," JMIR Mhealth Uhealth, vol. 7, no. 4, p. e11500, 2019, doi: 10.2196/11500.

[23] J. Brooke, "SUS: A 'Quick and Dirty' Usability Scale," Usability Evaluation In Industry, pp. 207-212, 1996, doi: 10.1201/9781498710411-35.

[24] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User Acceptance of Information Technology: Toward A Unified View1," MIS Quarterly, vol. 27, no. 3, pp. 425-478, 2003, doi: 10.2307/30036540.

[25] R. L. Spitzer, K. Kroenke, J. B. W. Williams, and B. Löwe, "A Brief Measure for Assessing Generalized Anxiety Disorder," Arch Intern Med, vol. 166, no. 10, p. 1092, 2006, doi: 10.1001/archinte.166.10.1092.

[26] P. A. Harris et al., "Research electronic data capture (REDCap)-A metadata-driven methodology and workflow process for providing translational research informatics support," Journal of Biomedical Informatics, vol. 42, no. 2, pp. 377-381, 2009, doi: 10.1016/j.jbi.2008.08.010.

[27] J. Torous, J. Nicholas, M. E. Larsen, J. Firth, and H. Christensen, "Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements," Evid Based Mental Health, vol. 21, no. 3, pp. 116-119, 2018, doi: 10.1136/eb-2018-102891.

[28] J. M. Isautier et al., "People's Experiences and Satisfaction With Telehealth During the COVID-19 Pandemic in Australia: Cross-Sectional Survey Study," J Med Internet Res, vol. 22, no. 12, p. e24531, 2020, doi: 10.2196/24531.

[29] A. Direito, Y. Jiang, R. Whittaker, and R. Maddison, "Apps for IMproving FITness and Increasing Physical Activity Among Young People: The AIMFIT Pragmatic Randomized Controlled Trial," J Med Internet Res, vol. 17, no. 8, p. e210, 2015, doi: 10.2196/jmir.4568.

[30] World Wide Web Consortium, "Web Content Accessibility Guidelines (WCAG) 2.2," W3C Recommendation, Oct. 2023. [Online]. Available: https://www.w3.org/TR/WCAG22/

[31] R. Harte, L. R. Quinlan, L. Glynn, A. Rodríguez-Molinero, et al., "Human-Centered Design Study: Enhancing the Usability of a Mobile Phone App in an Integrated Falls Risk Detection System for Use by Older Adult Users," JMIR mHealth and uHealth, vol. 5, no. 5, p. e71, 2017, doi: 10.2196/mhealth.7046.

[32] M. Chacornac, A. Faoro, J. Texereau, C. Billoet, et al., "Performance of an eHealth (NOMHAD) System Comprising Telemonitoring, Telenotification, and Telecoaching for Patients With Multimorbidity: Proof-of-Concept Study," JMIR Formative Research, vol. 6, no. 3, p. e32205, 2022, doi: 10.2196/32205.

[33] S. Zhang, X. Jia, Z. Zhao, and C. Zhou, "Design features of interfaces and acceptance of ePHR among elderly individuals in China: Insights from an expanded UTAUT framework," DIGITAL HEALTH, vol. 12, 2026, article 20552076261450778, doi: 10.1177/20552076261450778.

[34] X. Liu, B. Qiu, Y.-L. Ling, and K. Su, "Interface Design of Elderly mHealth Apps Based on A-Kano Model and Three-level Emotional Design Theory," Journal of Imaging Science and Technology, vol. 70, no. 1, pp. 1-11, 2026, doi: 10.2352/j.imagingsci.technol.2026.70.1.010409.

[35] B. Wang, J. Li, and Q. Zhu, "Structure Over Color: Eye-Tracking Evidence on Navigation Depth and Elderly Users' Online Trust in Health Apps," IEEE Access, vol. 14, pp. 5026-5036, 2026, doi: 10.1109/ACCESS.2026.3651334.

[36] S. Zhang and J. Song, "An empirical investigation into the preferences of the elderly for user interface design in personal electronic health record systems," Frontiers in Digital Health, vol. 5, 2024, article 1289904, doi: 10.3389/fdgth.2023.1289904.

[37] S. Choi and C. J. Chlebek, "Exploring mHealth design opportunities for blind and visually impaired older users," mHealth, vol. 10, p. 17, 2024, doi: 10.21037/mhealth-23-65.

[38] J. Qin, Z. Wu, H. Zhang, and X. Qi, "Designing age-friendly interfaces and wearable-environment integrations to enhance medication adherence: a discrete choice experiment among older adults," Frontiers in Medicine, vol. 13, 2026, article 1749507, doi: 10.3389/fmed.2026.1749507.

[39] K. Liu, Z. Sun, X. Ren, and D. Tao, "Effects of information framing cues and age on the comprehension of personal health records for self-care behaviors: An eye-tracking study," Journal of the American Medical Informatics Association, vol. 32, no. 7, pp. 1174-1185, 2025, doi: 10.1093/jamia/ocaf085.

[40] V. Balakrishnan, T. Kumaresan, H. Surendra, M. Allsop, et al., "Developing and assessing MyTBCompanion: A tri-lingual integrated video observed therapy app for tuberculosis patient management in Malaysia and Indonesia," PLOS One, vol. 20, no. 4, p. e0320394, 2025, doi: 10.1371/journal.pone.0320394.

[41] M. Liu, T. Jiang, K. Yu, S. Wu, M. Jordan-Marsh, and I. Chi, "Care Me Too, a Mobile App for Engaging Chinese Immigrant Caregivers in Self-Care: Qualitative Usability Study," JMIR Formative Research, vol. 4, no. 12, p. e20325, 2020, doi: 10.2196/20325.

[42] X. Wu and J. M. Beer, "Usability Evaluation of Telepresence Interfaces for Older Adults," Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 65, no. 1, pp. 591-595, 2021, doi: 10.1177/1071181321651172.

[43] U. Saxena and N. Sharma, "Elderly User Engagement with Mobile Health Applications: A Qualitative Analysis of User Interface and User Experience for Enhanced Technology Acceptance through a Scoping Review," Journal of Qualitative Research in Health Sciences, vol. 14, article 1346, 2025, doi: 10.34172/jqr.1346.

[44] M. Ghorbanian Zolbin and S. Kujala, "Evaluating the Usability of a Patient Portal: Insights from Older Migrant Adults," Studies in Health Technology and Informatics, vol. 336, pp. 2344-2348, 2026, doi: 10.3233/SHTI260689.