International Wound Journal

WILEY | IWJ

ORIGINAL ARTICLE OPEN ACCESS

Validation of a Clinical Decision-Support Algorithm for Chronic Wound Classification and Treatment: An Expert Consensus

Raquel Marques[1,2] | Carla Pais-Vieira[1,2] | Marcos Lopes[3] | João Neves-Amado[1,2] | Paulo Alves[1,2]

[1]Faculdade de Ciências da Saúde e Enfermagem, Centre for Interdisciplinary Research in Health, Universidade Católica Portuguesa, Porto, Portugal | [2]Faculdade de Ciências da Saúde e Enfermagem, School of Nursing Department, Universidade Católica Portuguesa, Porto, Portugal | [3]School of Nursing Department, Universidade Federal Ceará, Fortaleza, Brazil

Correspondence: Raquel Marques (rasilva@ucp.pt)

Received: 13 February 2026 | **Revised:** 4 May 2026 | **Accepted:** 11 May 2026

Keywords: clinical decision support system | consensus | diagnosis | observer variation | wounds and injuries

ABSTRACT

Accurate chronic wound classification is essential for appropriate management, yet diagnostic variability persists in routine practice. Transparent, rule-based decision-support tools may improve standardisation but require validation against expert judgement under clearly defined conditions. To evaluate inter-expert agreement, agreement between a rule-based algorithm and an expert-consensus reference standard, diagnostic accuracy as a complementary measure, exploratory comparison with a non-expert nurse, and expert agreement with algorithm-generated therapeutic recommendations. Thirty anonymised standardised clinical cases were classified by the algorithm and one non-expert nurse. Thirty wound-care experts, including 26 nurses, three physicians, and one researcher, were organised into six independent panels of five and classified case subsets, yielding 150 ratings. A consensus reference diagnosis was defined a priori as agreement by at least 3/5 experts. The primary outcome was algorithm–consensus agreement using Cohen's κ. Expert reliability was assessed using Krippendorff's α and Fleiss' κ. Recommendation agreement was dichotomised and analysed exploratorily. Expert agreement was low to moderate (Krippendorff's $\alpha = 0.26–0.60$), highest for pressure ulcers/injuries and venous leg ulcers, and lowest for mixed or unknown leg ulcers and diabetic foot ulcers. Consensus was reached in 29 of 30 cases. The algorithm achieved 86.2% accuracy (25/29) and substantial agreement ($\kappa = 0.70$, 95% CI 0.46–0.94). Nurse accuracy was 72.4% (21/29, $p = 0.219$). Experts endorsed 85.2% of therapeutic recommendations. The algorithm showed promising agreement under controlled conditions, supporting further prospective validation in larger, balanced real-world datasets.

1 | Introduction

Chronic wounds represent a major and growing public health challenge due to their prolonged healing trajectories, high recurrence rates, and substantial impact on quality of life, health-care utilisation, and costs [1]. Global prevalence estimates indicate that chronic wounds affect approximately 1–2 per 1000 individuals, disproportionately burdening older adults and persons with multimorbidity, vascular disease, diabetes, and reduced mobility. Among the most prevalent entities are pressure ulcers/injuries, venous leg ulcers, diabetic foot ulcers, and each requiring distinct diagnostic reasoning and therapeutic approaches [2].

Accurate classification of wound aetiology is a cornerstone of effective management, as diagnostic attribution directly informs key clinical decisions, including pressure redistribution, compression therapy, offloading strategies, vascular referral, and selection of dressings or advanced therapies. International

Key Points

- Chronic wound classification showed persistent inter-expert variability, especially in mixed and diabetic wound categories.
- The rule-based algorithm showed substantial agreement with expert consensus in a controlled validation setting.
- Further prospective validation is needed to assess performance, usability and clinical impact in routine care.

guidelines for pressure ulcers/injuries, venous leg ulcers, and diabetic foot disease consistently emphasise that structured assessment and correct classification are prerequisites for evidence-based treatment and improved outcomes [3–5]. Nevertheless, diagnostic classification in routine practice remains challenging, particularly in the presence of mixed aetiologies, atypical presentations, or incomplete clinical information.

Empirical evidence shows that interrater agreement in wound classification is often limited and highly dependent on rater expertise, clarity of definitions, and case complexity. Studies examining pressure ulcers/injuries staging and chronic wound categorisation report low to moderate agreement, even among trained professionals, with variability increasing in borderline or mixed cases [6, 7]. Prior work has highlighted substantial heterogeneity in diagnostic practices across settings, underscoring the need for tools that support consistency and reduce unwarranted variation [8]. This diagnostic variability not only affects treatment choices but may also contribute to delayed healing, inappropriate interventions, and inefficient use of healthcare resources [9].

In response to these challenges, clinical decision-support systems (CDSS) have been increasingly proposed to promote standardisation, guideline adherence, and safer clinical decision-making in wound care. Digital decision-support tools can synthesise clinical information and provide structured recommendations, potentially supporting guideline-concordant practice, particularly in community and long-term care settings where wound assessment is often undertaken by generalist nurses without formal specialist training and access to wound specialists may be limited [7, 10, 11]. The World Health Organisation has explicitly recommended the adoption of digital decision-support interventions as part of health system strengthening strategies, while emphasising the need for robust validation and governance of the underlying algorithms [12].

CDSS can broadly be divided into two categories: (i) knowledge-based, rule-driven systems that rely on explicit IF–THEN logic derived from clinical guidelines and expert consensus; and (ii) data-driven systems based on machine learning or deep learning approaches that infer patterns from large datasets [13]. While recent advances in Artificial Intelligence (AI) have enabled high-performance image-based wound classification and segmentation models [14], rule-based systems remain particularly relevant in nursing practice because of their transparency, interpretability, and closer alignment with clinical reasoning and guideline logic. These characteristics are critical for trust, explainability, and regulatory acceptability in clinical contexts [7]. In this respect, the rule-based nature of the system is not merely a technical choice but also a governance advantage, aligning with emerging regulatory frameworks such as the European Union AI Act, which emphasises transparency, human oversight, and traceability in clinical decision-support tools [15].

Several studies have demonstrated that well-designed rule-based algorithms can achieve substantial agreement with expert judgement in wound classification and management decisions. In some cases, agreement between algorithms and experts has exceeded 85%–90%, particularly for common wound types and clearly defined clinical scenarios [6, 16, 17]. However, much of the existing literature relies on content validation, simulated cases, or limited evaluation designs, and relatively few studies formally benchmark algorithmic outputs against an explicit expert-consensus reference standard. Moreover, validation often focuses on diagnostic labels alone, without examining agreement regarding therapeutic recommendations, which represent a distinct and clinically meaningful dimension of decision support. Beyond diagnostic accuracy, the evaluation of therapeutic recommendations is increasingly recognised as essential; a health professional may agree with a diagnostic label yet disagree with the proposed management, or vice versa [18]. Accordingly, agreement with therapeutic recommendations constitutes a complementary and clinically meaningful outcome when assessing the utility of decision-support systems.

A central methodological challenge in validating wound-classification algorithms is the absence of an objective gold standard. Histopathology or imaging rarely provide definitive labels for chronic wound aetiology, and diagnosis remains largely clinical. Consequently, many validation studies rely on expert panels to construct a reference standard, typically through majority agreement or consensus procedures [19]. This approach allows approximation of a clinically acceptable "ground truth" while simultaneously revealing the extent of expert disagreement. Importantly, documenting inter-expert variability is not a limitation but rather a key outcome, as it reflects the intrinsic ambiguity of real-world wound assessment and sets realistic performance expectations for decision-support tools.

Within this context, a research project led to the development of *CWS_Validation*, a rule-based clinical decision-support mobile application designed to assist healthcare professionals in the assessment and treatment of persons with chronic wounds. The application integrates multimodal clinical inputs and applies explicit IF–THEN decision rules derived from guidelines and expert knowledge to support wound classification and therapeutic planning. Prior to clinical deployment, its underlying algorithms underwent content validation involving a large panel of clinicians with varying levels of expertise. However, content validity alone is insufficient to establish clinical reliability, and further empirical validation against expert judgement is required.

The present study addresses this gap by conducting a structured validation of the diagnostic and therapeutic performance of the *CWS_Validation* algorithm using a curated set of anonymised clinical cases assessed by expert raters. Specifically, this study aimed to: (i) quantify inter-expert agreement in chronic wound

classification; (ii) evaluate agreement between a rule-based algorithm and an expert-consensus reference standard, with diagnostic accuracy as a complementary descriptive measure and exploratory comparison with a non-expert nurse; and (iii) assess expert endorsement of the algorithm's therapeutic recommendations, using an exploratory approach that accounted for the clustered structure of the data. Importantly, this study also addresses a frequently overlooked dimension in CDSS validation by explicitly evaluating agreement with algorithm-generated therapeutic recommendations, recognising that diagnostic concordance does not necessarily translate into agreement on therapeutic plans [18, 20]. This distinction aligns with contemporary perspectives on clinical reasoning under uncertainty, in which action-oriented decisions are often guided by precautionary and safety principles rather than strict etiological certainty [21].

2 | Materials and Methods

2.1 | Design Study

This study adopted a cross-sectional validation design to evaluate the diagnostic classification of a rule-based clinical decision-support algorithm for chronic wound assessment and expert agreement with its therapeutic recommendations, consistent with prior cross-sectional validation approaches used to validate wound care algorithms [6]. The study followed a structured expert-panel validation approach, commonly used when no objective gold standard is available [22].

A curated dataset of 30 anonymised clinical cases was used, each representing a distinct wound scenario. The study design was intended to assess inter-expert agreement, agreement between the algorithm and an expert-consensus reference standard, diagnostic accuracy as a complementary descriptive measure, exploratory comparison with a non-expert nurse, and expert agreement with algorithm-generated therapeutic recommendations under controlled conditions.

2.2 | Sample and Participants

Each clinical case was independently rated by five wound-care experts from the corresponding panel and by one non-expert nurse, yielding 150 expert ratings (5 × 30 cases) and 30 nurse ratings. To ensure feasibility and reduce cognitive burden, experts were organised into six independent panels, each composed of five experts and evaluating a mutually exclusive subset of five clinical cases. Within each panel, all five experts rated the same cases independently, enabling within-panel inter-rater agreement and a consensus-based reference classification. This design resulted in an incompletely crossed rater–case structure, in which not all raters evaluated all cases. Accordingly, inter-rater agreement estimates reflect within-panel reliability rather than agreement across a fully shared rater framework, and this was considered when interpreting reliability.

Experts were health professionals with recognised national expertise in wound care, affiliated with scientific societies and meeting predefined criteria (advanced training, ≥ 10 years' professional experience, and ≥ 5 years' clinical wound-care practice). Thirty experts participated (mean age 45.0 years, SD 6.79; 46.7% male), including 26 nurses, three physicians and one researcher, with a mean of 21 years' professional experience and 18 years in wound care, 70% had relevant publications and 80% held training roles. Non-expert nurses were recruited from seven healthcare units in mainland Portugal, spanning outpatient, home-care, rehabilitation and hospital settings, and providing care to adults and older patients with complex wounds.

2.3 | Procedures

Clinical cases were derived from data collected using the *CWS_Validation* app within a prospective multicentre cohort study conducted by the research team. Nurses followed individuals with complex wounds (pressure ulcers/injuries, leg ulcers and diabetic foot ulcers) across four assessment time points over approximately 4 weeks, recording structured clinical data and capturing wound images. For the present validation, baseline data and wound photographs taken after dressing removal and cleansing with normal saline solution or potable water were used to construct clinical cases.

Cases incorporated demographic and clinically relevant variables, including age, sex, comorbidities and associated conditions, body mass index, prior vascular surgery, wound recurrence and tobacco use in the preceding 6 months. Wound descriptors included anatomical location and key signs and symptoms (e.g., oedema, hyperpigmentation, atrophic blanche, telangiectasias, varicose veins, intermittent claudication, neuropathic and nociceptive pain, prior amputations etc.). Vascular assessment variables comprised posterior tibial and dorsalis pedis pulses, ankle–brachial index, Doppler signal and/or toe–brachial index. Whether the wound was located over a bony prominence or associated with a medical device was also recorded. All raters evaluated standardised case descriptions and clinical images under consistent presentation conditions. Assessments were performed independently, and raters were blinded to both the algorithm output and the evaluations of other raters.

Cases were reviewed independently via an electronic questionnaire (Qualtrics). Experts selected the wound type and rated agreement with the therapeutic plan based on local wound changes using a unipolar adjectival scale (1 = totally inappropriate; 2 = slightly appropriate; 3 = partially appropriate; 4 = very appropriate; 5 = totally appropriate; 6 = "don't know/prefer not to answer"), with optional free-text fields for comments.

Agreement with therapeutic recommendations was therefore assessed using an ordinal scale reflecting degrees of perceived appropriateness. For analytical purposes, responses were dichotomised to distinguish clear endorsement from non-endorsement. This approach was adopted because of the limited sample size and sparse distribution across ordinal categories, which precluded the reliable use of ordinal modelling approaches. Accordingly, the resulting binary variable should be interpreted as reflecting overall endorsement rather than gradations in expert agreement.

Because no objective external gold standard exists for chronic wound aetiology, a consensus-based reference diagnosis was

defined a priori as agreement by at least three of five experts on the same diagnostic category. This approach is consistent with established practice in diagnostic research relying on expert panels. However, the resulting reference classification should be interpreted as an operational approximation of expert judgment rather than a definitive diagnostic ground truth. The resulting consensus diagnosis was used as the reference standard for analyses of diagnostic agreement and algorithm performance, while cases not meeting this threshold were excluded from analyses requiring a reference diagnosis.

The evaluated clinical decision-support system is a guideline-informed, rule-based algorithm designed to support the classification and management of chronic wounds. It implements hierarchical decision logic reflecting guideline priorities and clinical reasoning pathways, applying rules sequentially to generate a primary wound classification and associated therapeutic recommendations. Examples of the diagnostic decision structure are provided in Appendix A. The treatment recommendation logic (Appendix B) incorporates variables such as the presence of slough and/or necrotic tissue, signs of infection, exudate level (none, low, moderate or high), and maceration of the wound edges and/or periwound skin.

The algorithm operates on a predefined set of structured clinical inputs routinely collected during wound assessment and generates a wound classification and associated treatment recommendations only when all required metadata for the relevant decision rules are complete; otherwise, diagnosis and therapeutic recommendations are assigned by the user.

2.4 | Ethical Considerations

This study was approved by the Health Ethics Committee of the Administração Regional de Saúde do Norte (Opinion No. CE/2023/58) and by the Unidade Local de Saúde da Guarda (Decision No. 55/2023). For the development of clinical cases, the investigators had exclusive access to pseudonymized data stored on servers located within the European Union, in accordance with the General Data Protection Regulation. Privacy, confidentiality and anonymity were rigorously ensured for participants, the collected data and the institutions involved. All participants provided informed written consent, including authorization for the capture and use of clinical images, as approved by the health ethics committees.

2.5 | Bias

Experts and the nurse were blinded to each other's ratings and to the algorithm output. All raters evaluated standardised case descriptions and clinical images under consistent presentation conditions, and assessments were performed independently. This approach aimed to minimise confirmation and observer bias, preserve the independence of assessments, and support comparability between human ratings and algorithm-generated classifications.

To reduce rater burden and support independent assessment, experts were organised into six independent panels, each composed of five experts and each assessing a subset of clinical cases. Case allocation was structured to include different wound types where possible; however, given the limited sample size and imbalanced case distribution, full balance across panels and wound categories could not be ensured. Accordingly, the possibility of residual case-mix or panel-composition bias cannot be excluded.

The use of curated clinical cases with complete structured information reduced missing-data bias for algorithm inputs and enabled controlled comparison across raters. However, this design may have reduced ecological variability compared with routine clinical practice, where documentation completeness, image quality, and contextual information are more heterogeneous.

2.6 | Statistical Methods

The primary outcome of this study was agreement between the algorithm and the expert-consensus reference standard, quantified using Cohen's κ. Diagnostic accuracy was calculated as a complementary descriptive measure. Descriptive and inferential analyses were undertaken with acknowledgement of the clustered structure of the data. Descriptive statistics summarised the distribution of wound types across algorithm classifications, nurse classifications and expert ratings. Categorical variables were reported as frequencies and percentages. Analyses of agreement with therapeutic recommendations were restricted to evaluable cases.

All clinical cases were curated to ensure completeness of the information required by the algorithm; therefore, no imputation procedures were necessary for algorithm inputs. At the level of expert evaluation, responses such as “not applicable” and “don't know/prefer not to answer” were treated as non-evaluable and excluded from analyses requiring definitive classification or agreement assessment, following a complete-case approach. Although this approach may introduce some bias if such responses are not randomly distributed, the frequency of non-evaluable responses was low.

Inter-rater reliability among experts was assessed using complementary measures for nominal categorical data: Krippendorff's alpha to quantify within-panel inter-rater reliability, and Fleiss' kappa to summarise category-level agreement across raters. These measures were used descriptively to characterise diagnostic consistency and agreement patterns. Given the limited number of cases per expert panel and the imbalance across diagnostic categories, wide confidence intervals for Krippendorff's alpha were expected. This behaviour is well documented in agreement studies with sparse and clustered data, where reliability estimates are highly sensitive to the number of items and category prevalence. Accordingly, wide confidence intervals should be interpreted primarily as a consequence of sample structure rather than as evidence of estimator instability or lack of methodological robustness [22, 23].

To explore classification patterns and related associations, contingency tables were constructed to compare: (i) algorithm classification versus the expert-consensus reference standard; (ii) nurse classification versus the expert-consensus reference

standard; and (iii) expert endorsement of algorithm-generated therapeutic recommendations. Paired comparison between algorithm and nurse classifications was performed using McNemar's exact test. Association between multicategory classifications was explored using χ^2 tests and Cramér's V, interpreted as measures of dependence rather than diagnostic performance. For binary outcomes and sparse contingency tables, Fisher's exact test was used in preference to Pearson's χ^2 test.

All contingency-based analyses were considered exploratory, given the limited sample size and the clustered structure of the data arising from multiple expert ratings per case.

All statistical analyses were performed in JASP (version 0.19.3.0) with R integration. Statistical significance was set at $p < 0.05$.

3 | Results

3.1 | Characteristics of the Evaluated Cases and Expert Ratings

A total of 30 clinical cases were included in the analysis. Each case was independently classified by the rule-based decision-support algorithm and by one non-expert nurse, yielding 30 algorithm-based and 30 nurse-based diagnostic classifications. In parallel, each case was evaluated by five independent expert raters, resulting in a total of 150 expert ratings.

Table 1 summarises the distribution of wound categories assigned by the algorithm, the nurse, and the expert raters. Across all three sources of classification, pressure ulcers/injuries were the most frequently assigned category. At the case level, the algorithm and the nurse each classified 66.7% of cases as pressure ulcers/injuries. The remaining classifications were distributed across a small number of venous leg ulcers, diabetic foot ulcers, mixed leg ulcers, and ulcers of unknown aetiology, with minor differences between the algorithm and nurse classifications in the less frequent categories.

TABLE 1 | Distribution of wound categories assigned by the algorithm, nurses, and expert raters.

Wound category	Algorithm (cases, $n = 30$)	Nurse (cases, $n = 30$)	Expert ratings ($n = 150$)
Pressure ulcers/injuries	20 (66.7)	20 (66.7)	105 (70.0)
Venous leg ulcer	3 (10.0)	2 (6.7)	21 (14.0)
Diabetic foot ulcer	3 (10.0)	2 (6.7)	8 (5.3)
Mixed leg ulcer	1 (3.3)	3 (10.0)	5 (3.3)
Unknown aetiology leg ulcer	3 (10.0)	1 (3.3)	4 (2.7)
Arterial ulcer	0 (0.0)	1 (3.3)	6 (4.0)
Device-related pressure ulcer/injuries	0 (0.0)	1 (3.3)	1 (0.7)
Total	30 (100)	30 (100)	150 (100)

Note: Frequencies and percentages are reported separately for case-level classifications (algorithm and nurse, $n = 30$ each) and for expert ratings ($n = 150$). Expert ratings correspond to five independent evaluations per case and therefore do not represent independent clinical observations.

At the expert-rating level, pooled classifications showed a similar overall pattern, with pressure ulcers/injuries remaining the predominant category, followed by venous leg ulcers and diabetic foot ulcers. Less frequent categories, including arterial ulcers, mixed leg ulcers, unknown aetiology leg ulcers, and device-related pressure ulcers/injuries, were assigned only occasionally.

Because each clinical case contributed multiple expert ratings, expert-level frequencies reflect the distribution of individual classifications rather than independent case-level observations. These data are therefore presented for descriptive purposes only and should be interpreted alongside the inter-rater agreement and consensus analyses reported in the following section.

3.2 | Agreement With Algorithmic Therapeutic Recommendations

Agreement between expert judgement and the algorithm's therapeutic recommendations was assessed at the level of individual expert ratings. Experts evaluated the clinical adequacy of the algorithm-generated therapeutic recommendation set using an ordinal scale. For analytical purposes, ratings indicating clear endorsement ("very appropriate" or "totally appropriate") were coded as agreement, whereas ratings reflecting partial or low adequacy ("totally inappropriate", "slightly appropriate" or "partially appropriate") were conservatively coded as disagreement. Responses marked as "don't know/prefer not to answer" were treated as non-evaluable and excluded from the relevant analyses. This conservative dichotomisation was adopted to capture unequivocal expert endorsement and may underestimate broader acceptability; accordingly, the resulting binary variable should be interpreted as reflecting clear endorsement rather than gradations in expert agreement.

Of the 150 total expert evaluations, 149 were valid and included in the analysis, with one response (0.7%) classified as "don't know" and excluded. Among the evaluable assessments, experts agreed with the algorithm-generated therapeutic recommendation in 127 cases (85.2%), while disagreement was observed in 22 cases (14.8%) (Table 2). The resulting binary agreement variable showed a proportion of agreement of 0.85 (SD = 0.36), indicating a generally high level of expert endorsement of the algorithm's therapeutic recommendations.

3.3 | Inter-Rater Agreement and Reference Classification

Inter-rater agreement was assessed within each group of five experts using Krippendorff's alpha at the nominal level. Across the six expert groups, α values ranged from 0.26 to 0.60, indicating low to moderate agreement (Table 3).

TABLE 2 | Agreement with algorithmic therapeutic recommendations.

Agreement with recommendation	*n*	Valid %
Disagree (0)	22	14.8
Agree (1)	127	85.2
Total	149	100.0

Note: Frequencies and valid percentages are calculated excluding "don't know" responses ($n = 149$).

TABLE 3 | Inter-rater agreement across expert groups.

Group	Krippendorff's α	95% CI	Interpretation
G1	0.46	−0.13–1.00	Moderate
G2	0.59	−0.08–0.76	Moderate
G3	0.49	0.18–1.00	Moderate
G4	0.60	−0.13–1.00	Moderate
G5	0.26	−0.03–1.00	Low
G6	0.54	−0.13–1.00	Moderate

Note: Krippendorff's alpha values reflect inter-rater reliability within each group of five experts evaluating five clinical cases.

Using a predefined majority rule, defined as agreement by at least three of five experts on the same diagnostic category, a consensus-based reference diagnosis was established for 29 of the 30 cases (96.7%).

These consensus labels were used as the operational reference standard for subsequent analyses of diagnostic agreement and for calculation of diagnostic accuracy as a complementary descriptive measure.

Agreement was higher for pressure ulcers/injuries and venous leg ulcers, whereas substantially lower agreement was observed for mixed leg ulcers, unknown aetiology leg ulcers, and diabetic foot ulcer categories. Category-level agreement assessed using Fleiss' κ supported this pattern, with moderate to high values for pressure ulcers/injuries ($\kappa = 0.54–0.84$) and venous leg ulcers ($\kappa = 0.63–0.78$), and markedly lower values for the remaining categories (Table 4).

Despite variability in inter-rater reliability, a consensus-based reference diagnosis could be established for most cases. An expert-consensus reference classification was defined a priori as agreement by at least three of five experts on the same diagnostic category. One case (Group 3, Case 13) did not meet the consensus threshold due to a dispersed pattern of expert ratings and was therefore excluded from analyses requiring a reference diagnosis.

A summary of inter-rater agreement and consensus achievement by group is presented in Table 3, and detailed case-level distributions of expert ratings and consensus outcomes are provided in Appendix C.

TABLE 4 | Category-level agreement (Fleiss' κ).

Category	κ range	Interpretation
Pressure ulcers/injuries	0.54–0.84	Moderate–high
Venous leg ulcers	0.63–0.78	Moderate–high
Mixed leg ulcer	−0.04–0.43	Low
Unknown aetiology leg ulcer	−0.09 to −0.04	Very low
Diabetic foot ulcer	0.19–0.25	Low

3.4 | Diagnostic Agreement Between the Algorithm and the Expert-Consensus Reference Standard

Diagnostic agreement between the algorithm and the expert-consensus reference standard was evaluated at the case level using the consensus-based classification derived from expert ratings. Only cases for which a reference diagnosis could be established were included in this analysis ($n = 29$).

The algorithm correctly classified 25 out of 29 cases with an expert-consensus reference diagnosis, corresponding to an overall accuracy of 86.2% within the evaluated case mix. Agreement beyond chance between the algorithm and the expert-consensus reference standard was substantial, with a Cohen's κ of 0.70 (95% CI: 0.46–0.94).

Cross-tabulation of algorithm predictions against the expert-consensus reference classification is presented in Table 5. Correct classifications were most frequently observed for pressure ulcers/injuries and venous leg ulcers, which were also the most prevalent categories in the dataset. Misclassifications occurred primarily in less frequent and diagnostically ambiguous categories, such as mixed aetiology leg ulcers and diabetic foot ulcers.

Overall, these findings indicate that the algorithm showed substantial agreement with the expert-consensus reference diagnosis, particularly for the most prevalent wound categories, suggesting potential utility for standardised wound classification within the evaluated case mix.

To compare diagnostic performance between the algorithm and non-expert nurse, a paired analysis was conducted using McNemar's exact test. This analysis included only cases with an established expert-consensus reference diagnosis ($n = 29$) and evaluated whether the algorithm or the nurse correctly classified each case relative to this operational reference standard.

Among the six discordant cases, the algorithm was correct while the nurse was incorrect in five cases, whereas the nurse was correct and the algorithm incorrect in one case. Although this asymmetry favoured the algorithm, the difference did not reach statistical significance (exact McNemar test, $p = 0.219$).

Overall, the algorithm showed numerically higher diagnostic accuracy than nurses (86.2% vs. 72.4%), but the limited number

TABLE 5 | Cross-tabulation between reference classification (gold standard) and algorithm predictions ($n = 29$).

Reference classification	Mixed	Venous	Diabetic	Pressure	Total
Mixed leg ulcer	1	0	0	0	1
Venous leg ulcer	0	3	0	2	5
Diabetic foot ulcer	0	0	1	0	1
Pressure ulcers/injuries	0	0	1	20	21
Total	1	3	2	22	29

Note: Reference classification was defined by expert consensus (≥ 3/5 raters). Only cases with an established reference diagnosis were included.

of discordant observations constrained statistical power. This accuracy estimate should be interpreted in light of over-representation of pressure ulcers/injuries in the case set, which are generally associated with higher inter-expert agreement and clearer diagnostic criteria than mixed ulcers or diabetic foot ulcers. These results should therefore be interpreted as exploratory and indicative rather than confirmatory.

To further characterise the relationship between algorithm-generated classifications and the expert-consensus reference standard, contingency analyses were performed at the case level ($n = 29$). The distribution of diagnostic categories assigned by the algorithm across reference categories is shown in Table 5.

Association measures showed a statistically significant relationship between the algorithm's predicted categories and the expert-consensus reference classification ($\chi^2 = 67.10$, df = 12, $p < 0.001$; Cramér's V = 0.88), indicating structural correspondence between classifications. These findings should be interpreted as evidence of categorical dependence, rather than as indicators of diagnostic performance.

Alignment was highest for pressure ulcers/injuries and venous leg ulcers, whereas discrepancies were more frequent in less prevalent and diagnostically ambiguous categories, such as mixed leg ulcers or diabetic ulcers. These patterns mirror the observed distribution of expert agreement across categories and reinforce the need to interpret performance estimates in light of the specific case mix.

3.5 | Agreement With Algorithmic Therapeutic Recommendations by Diagnostic Alignment

Agreement between expert judgement and the algorithm's therapeutic recommendations set was analysed at the level of individual expert ratings. Only evaluations for which a therapeutic recommendation was applicable and for which diagnostic alignment information was available were included, resulting in 149 evaluable observations. Overall, experts agreed with the algorithm-generated therapeutic recommendation set in 127 of 149 evaluations (85.2%), while disagreement was observed in 22 evaluations (14.8%). The resulting binary agreement variable had a mean of 0.85 (SD = 0.36), indicating overall clear endorsement rather than gradations in expert agreement.

Agreement was subsequently examined according to whether the algorithm's diagnostic suggestion was aligned with the expert-consensus classification. When diagnostic alignment was present, experts agreed with the therapeutic recommendation set in 107 of 125 evaluations (85.6%). When the algorithm's diagnostic suggestion was not aligned with expert consensus, agreement was observed in 20 of 24 evaluations (83.3%) (Table 6).

TABLE 6 | Agreement with algorithmic therapeutic recommendations by diagnostic alignment ($n = 149$).

Diagnostic alignment	Disagree *n* (%)	Agree *n* (%)	Total
Not aligned	4 (16.7)	20 (83.3)	24
Aligned	18 (14.4)	107 (85.6)	125
Total	22 (14.8)	127 (85.2)	149

Note: Percentages are row percentages. Only evaluations with applicable recommendations and available diagnostic alignment information were included.

To assess whether this numerical difference reflected a statistically meaningful association, a 2 × 2 contingency table was analysed. Given the sparse cell counts, Fisher's exact test was prioritised and did not indicate a statistically significant association between diagnostic alignment and agreement with the therapeutic recommendation set ($p = 0.757$). The chi-square test yielded a consistent result (χ^2 (1) = 0.082, $p = 0.774$). The estimated odds ratio was close to the null and imprecise, with wide confidensce intervals, supporting a cautious interpretation of this exploratory analysis.

Taken together, these descriptive and contingency-based results indicate that expert agreement with the algorithm-generated therapeutic recommendations set was consistently high, in the evaluated cases, regardless of whether the algorithm's diagnostic suggestion aligned with expert consensus. However, this finding should be interpreted cautiously given the exploratory nature of the analysis, the dichotomisation of ordinal ratings, and the limited variability in the outcome.

4 | Discussion

This study evaluated a rule-based clinical decision-support algorithm for chronic wound classification and treatment recommendations against an expert-consensus reference standard, while also quantifying diagnostic variability among experts and comparing algorithm performance with a non-expert nurse. Three main findings emerge. First, inter-expert agreement was

low to moderate across groups, with markedly higher consistency for pressure ulcers/injuries and venous leg ulcers than for mixed/unknown leg ulcers and diabetic categories. Second, despite this variability, an expert-consensus reference diagnosis could be established for nearly all cases, and the algorithm achieved substantial agreement with the expert-consensus standard and high overall accuracy. Third, experts showed high agreement with the algorithm's therapeutic recommendations, and this agreement did not differ meaningfully according to whether the algorithm's diagnostic suggestion aligned with expert consensus.

The interpretation of these findings was aligned with the predefined primary outcome, namely agreement between the algorithm and the expert-consensus reference standard quantified using Cohen's κ. Other analyses were considered complementary or exploratory. In particular, association measures, including χ^2 and Cramér's V, were interpreted as indicators of dependence between classifications rather than as measures of diagnostic performance, and non-significant findings were not overinterpreted.

Observed interrater reliability (Krippendorff's $\alpha = 0.26–0.60$) reinforces that chronic wound classification is frequently nontrivial and remains vulnerable to interpretive ambiguity, even among experienced clinicians. This finding is consistent with a substantial body of literature showing that wound classification depends on the integration of heterogeneous and partially subjective cues, such as comorbidities and risk factors, morphology, anatomical location, vascular assessment, mobility and that agreement decreases when cases are atypical, borderline, or of mixed aetiology [6–8].

Similar challenges have been reported in studies of pressure ulcer/injury staging, where inter-observer agreement among trained professionals is often only moderate [24, 25]. Establishing diagnostic consensus can also be challenging in leg ulcers and diabetic foot ulcers, given the complexity and overlap of clinical features [26, 27].

In the present study, category-level agreement was moderate to high for pressure ulcers/injuries and venous leg ulcers, but substantially lower for mixed leg ulcers, unknown leg ulcers, and diabetic foot ulcers categories. This pattern is coherent with clinical reality, as these latter entities often require integration of vascular assessment, neuropathy testing, and longitudinal information that may not be fully available in static case vignettes [4, 5]. Conceptually, these findings support the view that "diagnostic ground truth" in chronic wound care is probabilistic rather than absolute, particularly outside well-defined types. Similar arguments have been made in the broader diagnostic reasoning literature, which emphasises that disagreement among experts often reflects genuine uncertainty rather than error [20, 21].

Importantly, the high rate of consensus achievement in this study (96.7%) indicates that expert panels can still generate a pragmatic expert-consensus reference standard for validation purposes when no objective gold standard exists. This approach is widely accepted in observational and methodological research, particularly when expert judgement constitutes the de facto clinical reference [19, 28]. At the same time, the single case without consensus illustrates a critical implication for real-world decision support: algorithms should explicitly accommodate uncertainty rather than enforce deterministic outputs in situations where even experts disagree. Emerging guidance on trustworthy AI in healthcare similarly emphasises the importance of uncertainty-aware outputs and transparency about confidence limits [15, 29].

Against the expert-consensus reference, the algorithm achieved 86.2% accuracy and substantial agreement (Cohen's $\kappa = 0.70$), indicating clinically meaningful concordance. These findings align with previous studies showing that structured, rule-based wound algorithms can approximate expert judgement with relatively high fidelity, particularly for common wound categories and clearly specified clinical scenarios [6, 16, 17]. Comparable levels of agreement have also been reported in validation studies of algorithm-supported wound assessment tools and rule-driven CDSS in nursing contexts [10, 30].

The strong association between algorithm outputs and expert-consensus reference classifications (Cramér's $V = 0.88$) further suggests broad structural alignment in classification patterns within this dataset. At the same time, misclassifications were primarily observed in less frequent and more diagnostically complex categories, such as mixed aetiology and diabetic foot ulcers. This pattern suggests that performance was influenced not only by class imbalance but also by the inherent diagnostic ambiguity of these wound types. Similar challenges have been reported in machine-learning–based wound classifiers, in which performance typically degrades for underrepresented or heterogeneous classes [31, 32].

This convergence between expert disagreement and algorithmic error suggests that misclassification is not merely a technical limitation but may also reflect intrinsic clinical ambiguity. Consequently, algorithm refinement should focus not only on accuracy optimisation but also on mechanisms to detect uncertainty, request additional information, or flag cases for specialist review. Such approaches are increasingly advocated in the literature on safe and explainable clinical AI [33, 34].

Although the algorithm showed numerically higher accuracy than the nurse (86.2% vs. 72.4%), this difference did not reach statistical significance and should therefore be interpreted descriptively, particularly given the limited statistical power. Nonetheless, the direction of the effect is consistent with prior studies suggesting that decision-support tools may help standardise practice and reduce variability among non-specialist clinicians, particularly in community or resource-constrained settings [11]. Evidence from broader CDSS research indicates that such systems are most beneficial when they complement—not replace—clinical judgement, especially for less experienced users [7].

The present findings therefore align with a growing consensus that rule-based CDSS may function as cognitive scaffolding, supporting systematic reasoning and guideline adherence rather than outperforming clinicians outright. Future studies

including multiple nurses with varying experience levels will be necessary to quantify variability in human performance and to more robustly estimate comparative effectiveness.

An important consideration regarding the generalisability of these findings is that the participant sample was composed mainly of nurses. This reflects the organisation of wound care in our setting, where nurses have a central role in chronic wound assessment and management, and therefore supports the ecological validity of the study within similar healthcare systems. However, professional roles in wound assessment vary across countries and clinical contexts. In some settings, wound evaluation may be led more often by physicians, podiatrists, or multidisciplinary specialist teams. Consequently, the present findings may not be fully transferable to healthcare systems with different professional responsibilities, training pathways, or decision-making structures. Future studies should therefore include more internationally and professionally diverse samples to determine whether the observed levels of agreement and algorithm performance are maintained across different models of care.

Expert agreement with algorithmic therapeutic recommendations was high (85.2%), suggesting overall concordance with the rule-based guidance in the cases evaluated. However, this finding should be interpreted cautiously, as the analysis was exploratory and based on dichotomised agreement ratings, which may obscure variation in the degree of expert endorsement.

This finding is particularly important given that management decisions in chronic wound care span multiple therapeutic domains which are not reducible to diagnostic labels alone [4, 5].

The observed level of agreement should be interpreted in light of the method used to evaluate therapeutic recommendations. In this study, the original ordinal judgement scale was dichotomised into agreement versus disagreement in order to determine whether the overall set of recommendations generated by the algorithm was considered clinically acceptable. Although this approach facilitates interpretation, it entails a loss of information and does not capture gradations in the strength of expert endorsement. Such pragmatic dichotomisation strategies are commonly adopted in small-sample exploratory studies to ensure interpretability when ordinal modelling is not feasible. In the present study, this approach was therefore applied in the context of the limited sample size and sparse distribution across ordinal response categories. In addition, the recommendations were evaluated globally across multiple domains rather than as isolated therapeutic actions. This reflects the clinical logic underpinning the algorithm, in which specific recommendations are tailored to local wound characteristics and structured according to the concept of wound bed preparation. In routine practice, such recommendations are commonly considered as an integrated management package rather than as separate actions. However, this global approach may mask domain-specific disagreement, as different recommendation types may vary in their degree of consensus and clinical sensitivity. Future studies should therefore retain the ordinal response structure and evaluate recommendations separately across therapeutic domains, particularly in areas known to be more context-dependent or clinically controversial.

The absence of a statistically significant association between diagnostic correctness and agreement with therapeutic recommendations likely reflects the conservative and safety-oriented nature of many guideline-based actions, as well as the limited variability observed in the outcome.

This apparent dissociation may be explained by what can be termed a "precautionary convergence" effect, whereby clinicians tend to agree with management strategies that prioritise safety, monitoring, and risk mitigation, even in the presence of diagnostic uncertainty. Such behaviour is consistent with established models of clinical decision-making under uncertainty, in which therapeutic actions are often less dependent on precise etiological attribution than on anticipated risk and potential harm [20, 33]. Clinicians may reasonably accept recommendations that prioritise risk mitigation, monitoring, or referral even when diagnostic labels differ. This interpretation is consistent with conceptual models of clinical reasoning under uncertainty, in which action-oriented decisions are often guided by precautionary principles rather than strict etiological certainty [20].

Accordingly, the present findings should be interpreted as reflecting algorithm performance under standardised validation conditions, using structured clinical cases and an expert-consensus reference standard, rather than real-world effectiveness. This study represents a controlled validation phase within a staged evaluation framework for clinical decision-support systems, consistent with recent recommendations such as DECIDE-AI [35].

Overall, the findings support the potential value of transparent, rule-based decision-support systems in promoting standardisation and consistency in chronic wound assessment, particularly for prevalent and well-characterised wound types. This aligns with broader evidence that CDSS can enhance clinician performance when embedded within appropriate governance, usability, and training frameworks [7].

Compared with more complex AI models, the findings of this study suggest that rule-based approaches remain a valid, safe and comprehensible option, particularly in settings where interpretability and confidence in clinical decisions are critical [8]. Although AI-based models show promise, they face important limitations related to interpretability, reliance on large datasets, and validation in real-world clinical settings [14, 36]. This is not to claim superiority over data-driven AI, but rather to emphasise the value of knowledge-based systems as a robust foundation for developing hybrid algorithms grounded in validated clinical data and logically structured decision pathways. In this context, explainability features that explicitly link recommendations to guideline-derived rules can enhance trust, auditability and regulatory acceptability, representing key advantages over opaque machine-learning models.

Future research should involve larger, multi-centre datasets with balanced representation of wound aetiologies to enable more robust evaluation of algorithm performance across the full spectrum of chronic wound types, as well as the inclusion

of multiple non-expert clinicians. Future studies should also evaluate algorithm performance in prospective, real-world clinical settings to assess robustness, usability, and clinical impact under routine care conditions, including incomplete data, variable image quality, evolving clinical information, and workflow constraints. Beyond agreement metrics, future evaluations should examine downstream outcomes such as guideline adherence, healing trajectories, and patient-centred outcomes.

4.1 | Limitations

This study has several limitations. First, the sample size was relatively small ($n = 30$), and the distribution of wound types was imbalanced, with a predominance of pressure ulcers/injuries. The interpretation of diagnostic performance should therefore be considered in light of this case distribution. This imbalance may have influenced the observed diagnostic performance, as these categories, both in this dataset and in literature, tend to show higher inter-rater agreement. Consequently, overall accuracy and agreement estimates may be overestimated and should be interpreted as reflecting the specific case mix of the dataset rather than uniform performance across all wound types. This imbalance is also relevant for agreement statistics, as prevalence can influence the magnitude and interpretability of kappa-type coefficients and related measures in imbalanced datasets. Although prevalence-robust alternatives, such as Gwet's AC1 or prevalence-adjusted kappa, have been proposed, their application here was constrained by sparse counts in several diagnostic categories, which can render adjusted or class-specific estimates unstable and potentially misleading. Accordingly, conventional agreement metrics were prioritised, and findings should be interpreted with appropriate caution.

In addition, less frequent and more diagnostically complex wound categories, such as mixed leg ulcers and diabetic foot ulcers, were underrepresented and are inherently associated with greater diagnostic variability. Notably, misclassifications were concentrated in these less frequent and diagnostically complex categories, which also showed lower inter-expert agreement. This pattern suggests that reduced performance in these categories may reflect intrinsic diagnostic ambiguity rather than systematic algorithmic failure, consistent with prior reports of class imbalance and rare-category performance challenges in wound-care and medical AI research [14, 31].

Second, a further limitation relates to the panel-based study design. Experts were organised into six independent panels, each assessing a subset of cases, resulting in an incompletely crossed rater–case structure. While this design reduced rater burden and supported independent assessment within panels, it precluded estimation of agreement across a single unified expert panel. Consequently, inter-rater agreement estimates reflect within-panel consistency rather than overall agreement across all experts and may be influenced by panel-specific composition. This limits the comparability of agreement estimates across cases and constrains the generalisability of reliability metrics. The wide confidence intervals observed in agreement measures should also be interpreted in light of this study design. In particular, the limited sample size, the small number of cases per panel, the clustered rater–case structure, and the imbalance in category distribution may have affected the precision of agreement estimates. Accordingly, agreement metrics should be interpreted as reflecting overall patterns rather than precise point estimates. Future studies using fully or partially crossed designs would allow more robust estimation of global inter-rater reliability. In addition, variation in the devices and platforms used to view cases, such as monitor resolution and zoom functionality, may have influenced how clearly specific wound features were visualised, although comparison across cases is likely to have mitigated this bias.

Third, the use of curated clinical cases with complete and structured data represents a further limitation. While this approach enabled controlled comparison across raters and strengthened internal validity, it does not fully capture the variability, incompleteness, and contextual complexity of real-world clinical practice. In routine care, wound assessment is often influenced by heterogeneous data quality, incomplete documentation, variable image quality, and evolving clinical information, all of which may affect both human and algorithmic decision-making.

Fourth, only one non-expert nurse participated, limiting inference about variability among generalist health professionals.

Fifth, the definition of the expert-consensus reference standard also warrants consideration. In the absence of an objective gold standard for chronic wound aetiology, a consensus-based approach was used, representing an operational benchmark rather than a definitive ground truth. In the context of low to moderate inter-expert agreement, this approach may introduce uncertainty into the reference classification and, consequently, into the evaluation of algorithm performance. Importantly, this variability reflects the inherent complexity and subjectivity of chronic wound assessment, particularly in diagnostically ambiguous or mixed aetiologies, rather than solely a methodological constraint [19].

Sixth, the evaluation of therapeutic recommendations has important limitations. First, the dichotomisation of an ordinal scale simplified interpretation but resulted in loss of information, limiting the ability to capture gradations in the degree of expert endorsement. Second, recommendations were evaluated as integrated outputs rather than decomposed into specific therapeutic domains. While this reflects how decision-support tools are typically used in clinical practice, it limits the ability to identify domain-specific variability in agreement. This represents a trade-off between ecological validity and analytical granularity. In addition, the high observed agreement on therapeutic recommendations limited outcome variability, thereby restricting the strength of inferential conclusions.

Future studies should use larger and more balanced datasets, ideally via multicentre recruitment and/or stratified sampling, to support robust estimation of class-specific performance and the use of prevalence-robust agreement metrics. Fully or partially crossed rater designs would also allow more robust estimation of global inter-rater reliability. In addition, future evaluations should retain the ordinal structure of responses and evaluate agreement across specific therapeutic domains.

Such designs would also enable evaluation of calibration and decision-analytic measures that better capture clinical utility across heterogeneous wound types [28].

5 | Conclusion

In this expert-panel validation study, the rule-based algorithm showed substantial agreement with the expert-consensus reference diagnosis and favourable diagnostic accuracy within the evaluated case mix. Expert concordance with algorithmic therapeutic recommendations was also high, although this finding should be interpreted cautiously given the exploratory nature of the analysis. The study highlights persistent diagnostic ambiguity among experts, particularly for mixed and less common wound categories, underscoring both the potential value of decision support and the importance of explicitly managing uncertainty in algorithmic outputs. Overall, these findings provide preliminary empirical support for structured, transparent decision-support systems as tools to promote standardised, evidence-informed chronic wound care, while reinforcing the need for larger, prospective and more balanced validations before routine clinical implementation.

Acknowledgements

The authors thank the Centre for Interdisciplinary Research in Health. We would like to thank all the nurses who participated in the data collection and patients.

Funding

This work was supported by FCT—Fundação para a Ciência e a Tecnologia, I.P. by project reference UID/04279/2025 and DOI identifier https://doi.org/10.54499/UID/04279/2025—Centro de Investigação Interdisciplinar em Saúde.

Ethics Statement

This study was approved by the ethics committee for health of the Northern Regional Health Administration (Opinion No. CE/2023/58) and by the Local Health Unit of Guarda (Decision 55/2023). Raquel Marques Silva.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

1. M. Olsson, K. Järbrink, U. Divakar, et al., "The Humanistic and Economic Burden of Chronic Wounds: A Systematic Review," *Wound Repair and Regeneration* 27, no. 1 (2019): 114–125, https://doi.org/10.1111/wrr.12683.

2. L. Martinengo, M. Olsson, R. Bajpai, et al., "Prevalence of Chronic Wounds in the General Population: Systematic Review and Meta-Analysis of Observational Studies," *Annals of Epidemiology* 29 (2019): 8–15, https://doi.org/10.1016/j.annepidem.2018.10.005.

3. European Pressure Ulcer Advisory Panel, National Pressure Injury Advisory Panel, Pan Pacific Pressure Injury Alliance. *Prevention and Treatment of Pressure Ulcers/Injuries: Clinical Practice Guideline. The International Guideline*, 3rd ed., E. Haesler (EPUAP/NPIAP/PPPIA, 2019).

4. N. C. Schaper, J. J. van Netten, J. Apelqvist, et al., "Practical Guidelines on the Prevention and Management of Diabetes-Related Foot Disease (IWGDF 2023 Update)," *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy* 40, no. 3 (2024): e3657, https://doi.org/10.1002/dmrr.3657.

5. Wounds UK, *Best Practice Statement: Primary and Secondary Prevention in Lower Leg Wounds* (Wounds UK, 2024), https://wounds-uk.com/wp-content/uploads/2024/09/LR24_BPS_Prevention_WUK-web-v2.pdf.

6. J. M. Beitz and L. van Rijswijk, "A Cross-Sectional Study to Validate Wound Care Algorithms for Use by Registered Nurses," *Ostomy/Wound Management* 56, no. 4 (2010): 46–59.

7. C. Thompson, T. Mebrahtu, S. Skyrme, et al., "The Effects of Computerised Decision Support Systems on Nursing and Allied Health Professional Performance and Patient Outcomes: A Systematic Review and User Contextualisation," *Health Soc Care Deliv Res* 12, no. 40 (2024): 1–94, https://doi.org/10.3310/grnm5147.

8. M. G. Rippon, L. Fleming, T. Chen, A. A. Rogers, and K. Ousey, "Artificial Intelligence in Wound Care: Diagnosis, Assessment and Treatment of Hard-To-Heal Wounds: A Narrative Review," *Journal of Wound Care* 33, no. 4 (2024): 229–242, https://doi.org/10.12968/jowc.2024.33.4.229.

9. H. T. Mohammed, S. Bestavros, S. Mohsen, et al., "Assessing Clinician Consistency in Wound Tissue Classification and the Value of AI-Assisted Quantification: A Cross-Sectional Study," *International Wound Journal* 22, no. 6 (2025): e70691, https://doi.org/10.1111/iwj.70691.

10. S. Jordan, J. McSwiggan, J. Parker, G. A. Halas, and M. Friesen, "An mHealth App for Decision-Making Support in Wound Dressing Selection (WounDS): Protocol for a User-Centered Feasibility Study," *JMIR Research Protocols* 7, no. 4 (2018): e108, https://doi.org/10.2196/resprot.9116.

11. E. M. Quinn, M. A. Corrigan, J. O'Mullane, et al., "Clinical Unity and Community Empowerment: The Use of Smartphone Technology to Empower Community Management of Chronic Venous Ulcers Through the Support of a Tertiary Unit," *PLoS One* 8, no. 11 (2013): e78786, https://doi.org/10.1371/journal.pone.0078786.

12. World Health Organization, "WHO Guideline: Recommendations on Digital Interventions for Health System Strengthening," (2019), https://iris.who.int/server/api/core/bitstreams/c3c53f30-23cc-48d0-a3b5-c05ddd7f5349/content.

13. K. Kawamoto, C. A. Houlihan, E. A. Balas, and D. F. Lobach, "Improving Clinical Practice Using Clinical Decision Support Systems: A Systematic Review of Trials to Identify Features Critical to Success," *BMJ* 330 (2005): 765, https://doi.org/10.1136/bmj.38398.500764.8F.

14. D. M. Anisuzzaman, C. Wang, B. Rostami, S. Gopalakrishnan, J. Niezgoda, and Z. Yu, "Image-Based Artificial Intelligence in Wound Assessment: A Systematic Review," *Adv Wound Care (New Rochelle)* 11, no. 12 (2022): 687–709, https://doi.org/10.1089/wound.2021.0091.

15. European Parliament and Council, "Artificial Intelligence Act. In: Official Journal of the European Union," (2024), https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689.

16. J. M. Beitz and L. van Rijswijk, "Development and Validation of an Online Interactive, Multimedia Wound Care Algorithms Program," *Journal of Wound, Ostomy, and Continence Nursing* 39, no. 1 (2012): 23–34, https://doi.org/10.1097/WON.0b013e3182383f07.

17. P. C. B. Khong, L. N. Lee, and A. I. Dawang, "Modeling the Construct of an Expert Evidence-Adaptive Knowledge Base for a Pressure Injury

Clinical Decision Support System," *Inform* 4, no. 3 (2017): 20, https://doi.org/10.3390/informatics4030020.

18. M. Omar, S. Soffer, R. Agbareia, et al., "Sociodemographic Biases in Medical Decision Making by Large Language Models," *Nature Medicine* 31 (2025): 1873–1881, https://doi.org/10.1038/s41591-025-03626-6.

19. E. von Elm, D. G. Altman, M. Egger, et al., "The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies," *Journal of Clinical Epidemiology* 61, no. 4 (2008): 344–349, https://doi.org/10.1016/j.jclinepi.2007.11.008.

20. H. N. Garb, "Clinical Judgment and Decision Making," *Annual Review of Clinical Psychology* 1 (2005): 67–89, https://doi.org/10.1146/annurev.clinpsy.1.102803.143810.

21. R. Fahed, T. E. Darsaut, B. Farzin, M. Chagnon, and J. Raymond, "Measuring Clinical Uncertainty and Equipoise by Applying the Agreement Study Methodology to Patient Management Decisions," *BMC Medical Research Methodology* 20, no. 1 (2020): 214, https://doi.org/10.1186/s12874-020-01095-8.

22. L. C. M. Bertens, B. D. L. Broekhuizen, C. A. Naaktgeboren, et al., "Use of Expert Panels to Define the Reference Standard in Diagnostic Research: A Systematic Review of Published Methods and Reporting," *PLoS Medicine* 10, no. 10 (2013): e1001531, https://doi.org/10.1371/journal.pmed.1001531.

23. k. krippendorff, *Computing Krippendorff's Alpha-Reliability*. (University of Pennsylvania, 2011), https://repository.upenn.edu/server/api/core/bitstreams/0421f871-f005-4322-b06a-a66bec328e3b/content.

24. D. Beeckman, L. Schoonhoven, J. Fletcher, et al., "EPUAP Classification System for Pressure Ulcers: European Reliability Study," *Journal of Advanced Nursing* 60, no. 6 (2007): 682–691, https://doi.org/10.1111/j.1365-2648.2007.04474.x.

25. J. Stausberg, N. Lehmann, K. Kröger, I. Maier, and W. Niebel, "Reliability and Validity of Pressure Ulcer Diagnosis and Grading: An Image-Based Survey," *International Journal of Nursing Studies* 44, no. 8 (2007): 1316–1323, https://doi.org/10.1016/j.ijnurstu.2006.06.006.

26. M. Nou Howaldt, A. Diard, M. C. Courtois, S. Mestre, and J. Guillaumat, "Diagnostic Assessment for Venous Leg Ulcers: Recommendations From the Wounds and Healing Working Group of the French Society of Vascular Medicine," *Vascular Diseases* 50, no. 3 (2025): 127–132, https://doi.org/10.1016/j.vasdi.2025.04.001.

27. T. B. Santema, E. A. Lenselink, R. Balm, and D. T. Ubbink, "Comparing the Meggitt-Wagner and the University of Texas Wound Classification Systems for Diabetic Foot Ulcers: Inter-Observer Analyses," *International Wound Journal* 13, no. 6 (2016): 1137–1141, https://doi.org/10.1111/iwj.12429.

28. E. W. Steyerberg, A. J. Vickers, N. R. Cook, et al., "Assessing the Performance of Prediction Models: A Framework for Traditional and Novel Measures," *Epidemiology* 21, no. 1 (2010): 128–138, https://doi.org/10.1097/EDE.0b013e3181c30fb2.

29. World Health Organization, "Ethics and Governance of Artificial Intelligence for Health: WHO Guidance," (2021), https://iris.who.int/server/api/core/bitstreams/f780d926-4ae3-42ce-a6d6-e898a5562621/content.

30. J. Gagnon, J. Chartrand, S. Probst, et al., "Co-Creation and Evaluation of an Algorithm for the Development of a Mobile Application for Wound Care Among New Graduate Nurses: A Mixed Methods Study," *International Wound Journal* 21, no. 10 (2024): e70064, https://doi.org/10.1111/iwj.70064.

31. B. Rostami, D. M. Anisuzzaman, C. Wang, S. Gopalakrishnan, J. Niezgoda, and Z. Yu, "Multiclass Wound Image Classification Using an Ensemble Deep CNN-Based Classifier," *Computers in Biology and Medicine* 134 (2021): 104536, https://doi.org/10.1016/j.compbiomed.2021.104536.

32. C. Wang, D. M. Anisuzzaman, V. Williamson, et al., "Fully Automatic Wound Segmentation With Deep Convolutional Neural Networks," *Scientific Reports* 10, no. 1 (2020): 21897, https://doi.org/10.1038/s41598-020-78799-w.

33. B. Kompa, J. Snoek, and A. L. Beam, "Second Opinion Needed: Communicating Uncertainty in Medical Machine Learning," *npj Digital Medicine* 4, no. 1 (2021): 4, https://doi.org/10.1038/s41746-020-00367-3.

34. A. J. London, "Artificial Intelligence and Black-Box Medical Decisions: Accuracy Versus Explainability," *Hastings Center Report* 49, no. 1 (2019): 15–21, https://doi.org/10.1002/hast.973.

35. B. Vasey, M. Nagendran, B. Campbell, et al., "Reporting Guideline for the Early Stage Clinical Evaluation of Decision Support Systems Driven by Artificial Intelligence: DECIDE-AI," *BMJ (Clinical Research Ed.)* 377 (2022): e070904, https://doi.org/10.1136/bmj-2022-070904.

36. O. C. Onuh, H. T. Brydges, H. Nasr, E. Savage, S. Gorenstein, and E. Chiu, "Capturing Essentials in Wound Photography Past, Present, and Future: A Proposed Algorithm for Standardization," *Advances in Skin & Wound Care* 35, no. 9 (2022): 483–492, https://doi.org/10.1097/01.ASW.0000852564.21370.a4.

Appendix A

Diagnostic Decision Structure

Example Decision Rule for Wound Typology Suggestion (Venous Leg ulcer)

Section	Logical requirement	Operational definition (structured inputs)
Prerequisite	Data completeness	The algorithm generates a wound type suggestion only when all required metadata for the relevant decision rules are complete; otherwise, no automated classification is provided and the wound type is assigned by the user.
1. Venous disease	Peripheral venous disease = yes	Peripheral venous disease recorded as present.
2. Anatomical location	Location within lower-leg region	Upper third of leg OR middle third OR lower third OR medial malleolar region OR calf region OR lateral malleolar region (including laterality: left/right).
3. Venous indicators/ history	At least one venous feature present	Previous vascular surgery OR recurrent wound OR hyperpigmentation OR atrophie blanche OR oedema OR telangiectasias/varicose veins OR venous eczema.
4. Adequate arterial perfusion	Any one criterion indicating adequate perfusion	Posterior tibial and dorsalis pedis pulses present bilaterally OR Doppler waveform bi−/tri-phasic bilaterally OR ABPI 0.91–1.30 OR TBI > 0.70.
THEN (output)	Wound type	Venous leg ulcer.

Appendix B

Decision Logic for Treatment Recommendations

The treatment recommendation logic incorporates the presence of slough and/or necrotic tissue, presence of infection, exudate level (none, low, moderate or high), and maceration of the wound edges and/or periwound skin.

IF–THEN examples (three representative rules).

IF necrotic/slough tissue = *absent* AND infection = *absent* AND exudate = *none* AND periwound/edge maceration = *absent*.

THEN if debris is present, cleanse the wound and periwound skin with potable water or 0.9% normal saline at body temperature, protect skin integrity from damage.

IF necrotic/slough tissue = *present* AND infection = *absent* AND exudate = *high* AND periwound/edge maceration = *present*.

THEN cleanse the wound and periwound skin with potable water or 0.9% normal saline at body temperature, consider using an antiseptic (3) according to local protocol, debride (2) non-viable tissue, apply an atraumatic dressing with high absorbency (6), apply a skin protectant/barrier product (7) to the wound edges and periwound skin if necessary, increase dressing change frequency.

IF necrotic/slough tissue = *absent* AND infection = *present* AND exudate = *high* AND periwound/edge maceration = *present*.

THEN cleanse the wound and periwound skin, use an antiseptic (3) according to local protocol and perform ongoing/regular debridement (2) (to disrupt the biofilm and reduce microbial load), apply a dressing with topical antimicrobial activity (4) and high absorbency (6), apply a skin protectant/barrier product (7) to the wound edges and periwound skin, if necessary, increase dressing change frequency. Numeric markers (e.g., (2), (3), (4)) correspond to internal dressing category codes in the app's recommendation library.

Appendix C

Case-Level Distribution of Expert Ratings and Consensus Classification

Group	Case	Distribution of ratings (*n*)	Consensus category	Consensus achieved
G1	C1	Venous leg ulcer (3), Mixed leg ulcer (1), Unknown aetiology leg ulcer (1)	Venous leg ulcer	Yes
G1	C2	Pressure ulcer/injury (5)	Pressure ulcer/injury	Yes
G1	C3	Pressure ulcer/injury (5)	Pressure ulcer/injury	Yes
G1	C4	Pressure ulcer/injury (4), Device-related pressure ulcer (1)	Pressure ulcer/injury	Yes
G1	C5	Pressure ulcer/injury (5)	Pressure ulcer/injury	Yes
G2	C6	Pressure ulcer/injury (4), Arterial ulcer (1)	Pressure ulcer/injury	Yes
G2	C7	Pressure ulcer/injury (5)	Pressure ulcer/injury	Yes

Group	Case	Distribution of ratings (*n*)	Consensus category	Consensus achieved
G2	C8	Venous leg ulcer (4), Unknown aetiology leg ulcer (1)	Venous leg ulcer	Yes
G2	C9	Venous leg ulcer (4), Unknown aetiology leg ulcer (1)	Venous leg ulcer	Yes
G2	C10	Pressure ulcer/injury (5)	Pressure ulcer/injury	Yes
G3	C11	Mixed leg ulcer (3), Arterial ulcer (2)	Mixed leg ulcer	Yes
G3	C12	Pressure ulcer/injury (5)	Pressure ulcer/injury	Yes
G3	C13	Diabetic foot ulcer (2), Pressure ulcer/injury (1), Arterial ulcer (2)	Pressure ulcer/injury	No
G3	C14	Pressure ulcer/injury (5)	Pressure ulcer/injury	Yes
G3	C15	Pressure ulcer/injury (5)	Pressure ulcer/injury	Yes
G4	C16	Pressure ulcer/injury (4), Diabetic foot ulcer (1)	Pressure ulcer/injury	Yes
G4	C17	Pressure ulcer/injury (5)	Pressure ulcer/injury	Yes
G4	C18	Venous leg ulcer (4), Arterial ulcer (1)	Venous leg ulcer	Yes
G4	C19	Pressure ulcer/injury (5)	Pressure ulcer/injury	Yes
G4	C20	Pressure ulcer/injury (5)	Pressure ulcer/injury	Yes
G5	C21	Pressure ulcer/injury (5)	Pressure ulcer/injury	Yes
G5	C22	Pressure ulcer/injury (5)	Pressure ulcer/injury	Yes
G5	C23	Pressure ulcer/injury (5)	Pressure ulcer/injury	Yes
G5	C24	Diabetic foot ulcer (3), Pressure ulcer/injury (2)	Diabetic foot ulcer	Yes
G5	C25	Pressure ulcer/injury (3), Diabetic foot ulcer (2)	Pressure ulcer/injury	Yes
G6	C26	Venous leg ulcer (5)	Venous leg ulcer	Yes
G6	C27	Pressure ulcer/injury (5)	Pressure ulcer/injury	Yes
G6	C28	Pressure ulcer/injury (5)	Pressure ulcer/injury	Yes
G6	C29	Pressure ulcer/injury (3), Mixed leg ulcer (1), Unknown aetiology leg ulcer (1)	Pressure ulcer/injury	Yes
G6	C30	Pressure ulcer/injury (4), Venous leg ulcer (1)	Pressure ulcer/injury	Yes

Note: A consensus diagnosis was defined a priori as agreement by at least three out of 5 raters on the same diagnostic category. Cases not meeting this criterion were classified as having no consensus and were excluded from analyses requiring a reference diagnosis.