Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Rule-Based Bilingual Decision-Support Prototype for Foot-and-Ankle Orthosis Selection Knowledge Representation and Preliminary Face-Validity Assessment

**Mohamad Firas Wahbeh**

firaswahbeh76@gmail.com

Damascus University https://orcid.org/0009-0006-7589-0169

**Method Article**

**Keywords:** clinical decision support systems, ankle-foot orthosis, rule-based reasoning, orthotic prescription, digital health, Software as a Medical Device, face validity, sensitivity analysis

**Posted Date:** August 22nd, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-10645470/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** The authors declare no competing interests.

## Abstract

# Background

The selection of a foot-and-ankle orthosis draws on evidence spread across international consensus reports, institutional atlases, and heterogeneous systematic reviews, with no open digital tool consolidating this evidence into a single auditable instrument.

# Objective

To describe the knowledge representation, scoring architecture, and software implementation of Orthotica, a transparent, rule-based, offline-capable, bilingual (English/Arabic) decision-support prototype, and to formally characterise its evidentiary status through a preliminary face-validity assessment.

# Methods

We describe a four-layer knowledge base with additive rule weights and hard-exclusion flags, plus an insole-specific biomechanical module. We conducted a ± 20% weight-sensitivity analysis across five cases, a retrospective application to 30 de-identified cases including concordance with the original orthotist's prescription, and a structured review by ten multidisciplinary experts using a standardised instrument. All rule weights are disclosed in Supplementary Appendix 1.

# Results

Retrospective suitability scores ranged 76–97% (mean 89.5%) and matched the original prescription's device category in 30/30 cases and exact subtype in 28/30 (93%) concordance that partly reflects shared provenance between the rule base and the source clinic rather than independent validation. Weight perturbation preserved the leading recommendation in 80.2% of trials (995/1,240), ranging from 98% (category) to 71% (AFO subtype). Ten experts, drawn mainly from one region, rated the tool a mean 8.9/10 for utility (ICC = 0.72, 95% CI 0.51–0.89).

# Conclusion

Orthotica shows preliminary face validity as a transparent educational prototype. Its weights are author-assigned and unvalidated against prospective, blinded, independent outcomes and must not substitute for clinical judgment. A concrete Delphi and prospective-validation protocol is proposed.

## 1. INTRODUCTION

Prescribing a foot or ankle orthosis is the terminal step of judgments concerning muscle tone, manual muscle strength, joint stability, the flexibility versus fixation of any deformity, and functional demand [1–4]. At least five nested device categories coexist within foot-and-ankle bracing foot orthosis, supramalleolar orthosis, dynamic ankle-foot orthosis (AFO), ground-reaction AFO, and the broader AFO family, which alone contains nine distinguishable subtypes and eleven ankle-joint mechanisms [4,5].

Matching a patient to the correct point in this taxonomy draws on evidence dispersed across international consensus reports [1,2], systematic reviews of variable methodological quality [5–7], and non-peer-reviewed manufacturer guidance. A 2024 systematic review restricted to randomised trials in children with cerebral palsy identified only eleven eligible trials, all at high risk of bias, despite orthotic bracing being one of the commonest interventions in this population [7]. No openly available digital instrument currently consolidates this heterogeneous evidence into a single, auditable decision-support tool.

Clinical decision support systems (CDSS) have been used since the 1980s to structure complex medical judgments, with documented benefits for consistency alongside well-described risks — alert fatigue, automation bias, and harm when recommendations are trusted beyond their evidentiary basis [8]. These risks are amplified, not reduced, when a CDSS is a heuristic prototype rather than a formally validated instrument, making transparent design and explicit disclosure of limitations a matter of basic scientific integrity.

This paper describes Orthotica, a bilingual, offline-capable web application built as an engineering and methodological prototype. Our contribution is deliberately narrow: we report the knowledge representation, scoring architecture, and software implementation, and we formally characterise the system's current evidentiary status through a preliminary face-validity assessment using retrospective cases, prescription concordance, and structured expert feedback. We explicitly do not claim any measure of diagnostic or prognostic accuracy derived from prospective outcome data. The complete rule base is provided as Supplementary Appendix 1 for full reproducibility.

## 2. BACKGROUND AND RATIONALE

The ISO terminology adopted in standard institutional atlases classifies lower-limb orthoses by the joints they span, but clinically meaningful distinctions are functional: a supramalleolar orthosis controls coronal-plane subtalar motion without controlling sagittal-plane ankle motion, a full AFO offers the highest degree of mechanical control, and a ground-reaction design substitutes an anterior for a posterior shell to generate an external knee-extension moment during stance, most often discussed in crouch gait in cerebral palsy [3–5].

Much of the empirical AFO literature has developed within cerebral palsy populations, reflecting the historical role of International Society for Prosthetics and Orthotics (ISPO) consensus conferences on cerebral palsy [1] and poliomyelitis [2]. Systematic reviews repeatedly note substantial heterogeneity in how orthoses are described and reported, limiting quantitative synthesis and direct comparability across

studies [5–7]. This combination of a large descriptive literature and a comparatively thin base of high-quality comparative trials is precisely the situation in which a transparent, structured decision aid can organise existing knowledge and equally the situation in which such an aid must not be mistaken for a validated clinical algorithm.

A structured tool can also encode standardised clinical measures as discrete, inspectable inputs, each with its own reliability profile. The Gross Motor Function Classification System was developed through formal Delphi-type consensus among paediatric rehabilitation experts and has demonstrated acceptable inter-rater reliability [9]. The Modified Ashworth Scale's original inter-rater testing was restricted to elbow flexors in thirty patients, with reliability for other muscle groups reported inconsistently since [10]. The Foot Posture Index has documented construct validity and generally acceptable reliability in adults and children [11, 12]. The navicular drop test uses an original diagnostic threshold of approximately ten millimetres for excessive pronation [13]. The Silfverskiöld test for isolated gastrocnemius tightness has shown only poor-to-moderate inter- and intra-rater reliability in recent dedicated reliability studies [14]. Safe incorporation of such tests into a CDSS requires explicit acknowledgement of their published reliability so that the system's noise floor is transparent to users [8].

## 3. SYSTEM DESCRIPTION

# 3.1 Design Philosophy and Knowledge Base

Orthotica was designed around three priorities: transparency, so every contributing factor behind a recommendation is inspectable; conservatism, so the tool flags for referral rather than forcing a recommendation when a contraindication is present; and accessibility, so it is free, fully offline-capable, and bilingual in English and Arabic. A rule-based architecture was chosen over a statistical or machine-learning approach specifically because every scoring decision can be traced to a named, citable clinical rationale a precondition for the external audit this paper invites [8]. This choice trades any claim to empirically optimised performance for complete interpretability.

Figure 1 shows the application's entry screen, exposing the four-step clinical workflow (Home → Assessment → Results → Library → Sources), the bilingual language toggle with right-to-left Arabic rendering, and direct access to the curated evidence library.

The knowledge base comprises four linked layers. Layer 1 scores five top-level device categories. Layer 2, activated when the AFO family leads, scores nine named subtypes (solid/rigid, posterior leaf-spring, articulated free-motion, plantarflexion-stop, carbon-fibre dynamic-response, tone-reducing, dorsiflexion-assist/night-stretch, soft fabric-reinforced, and patellar tendon-bearing designs). Layer 3, activated for typically articulated devices, scores eleven ankle-joint mechanisms. Layer 4 is the insole-specific biomechanical refinement module described below.

Each scoreable item carries discrete rules a single testable clinical statement (e.g., ‘moderate-to-severe spasticity’) paired with a signed weight, with a hard-exclusion flag where the literature describes an

outright contraindication. Rule content was developed by the author from the consensus and review literature cited above, generally available manufacturer clinical guidance, and informal iterative discussion with practising clinicians during development. No structured survey instrument was administered at that stage and no inter-rater agreement statistic was computed for the weights: the knowledge base is therefore a literature-grounded, practitioner-informed first draft, not the output of a formal Delphi exercise. This is the single most important methodological characteristic of the present work and is reiterated in § 5.3, § 8, and § 9. The complete rule base, including every weight, exclusion flag, and its source citation, is provided as Supplementary Appendix 1.

## 3.2 Scoring Architecture and Insole-Specific Refinement

Each candidate item begins from a neutral baseline; matched positive-indication rules add weight, matched caution rules subtract weight, and matched hard-exclusion rules subtract a large fixed penalty and set an exclusion flag. Ranking uses the unclamped raw sum, while a clamped version (bounded 2–98%) is displayed as a ‘suitability match’ for interpretability and is explicitly not a calibrated probability of clinical success. Once the leading category is set, the pipeline cascades through Layers 2–3 as applicable; at each layer the top-ranked item is the leading recommendation, the next four are ranked alternatives, and every fired rule is shown in an expandable rationale panel.

When foot orthosis is the leading category, a dedicated panel accepts the FPI-6 score [11, 12], navicular drop in millimetres [13], the result of Jack's test, and the result of the Silfverskiöld test [14]. These feed a secondary layer ranking four insole subtypes (rigid corrective shell, semi-rigid with deep heel cup, soft accommodative, and cushioned with lateral forefoot wedge) and independently generate a plain-language note when the Silfverskiöld result suggests isolated gastrocnemius tightness, since this finding is not itself correctable by any insole and instead points to a stretching program or referral. A non-reforming result on Jack's test functions as a hard exclusion for the rigid and semi-rigid subtypes, reflecting the clinical distinction between a correctable flexible flatfoot and a structurally fixed one.

## 3.3 Software Architecture and Regulatory Positioning

Orthotica is implemented as a client-side Progressive Web Application (HTML, CSS, vanilla JavaScript) with no server-side component and no transmission or persistent storage of clinical information; all computation occurs locally in the user's browser. The primary distributable is a single self-contained HTML file that can be opened offline; a parallel modular build with a service worker supports installation as an offline-capable application. Every clinical input field carries inline explanatory text, and a client-side export function generates a PDF summary of the assessment and rationale.

Under the International Medical Device Regulators Forum (IMDRF) risk-based framework for Software as a Medical Device (SaMD) [15], the present prototype supports rather than drives clinical decision-making and does not acquire or analyse images or signals from medical devices, placing it in the lowest SaMD risk category. Every output is accompanied by an inspectable rationale panel, and every contraindication flag is designed to trigger referral rather than silent override. Users are warned, in both interface languages, that rule weights have not been calibrated against prospective outcomes and that outputs

must be interpreted within full clinical context by qualified practitioners a transparency-first design intended to mitigate automation bias [8]. A running instance is freely accessible at https://superlativesmakager-116949.netlify.app/, requiring no account and collecting no personal data or analytics.

# 4. PRELIMINARY FACE-VALIDITY ASSESSMENT

This section formally characterises Orthotica's current evidentiary status through two parallel tracks retrospective application to de-identified clinical cases, including concordance with the original prescriber, and structured expert review presented as a preliminary face-validity assessment, not as clinical validation. ‘Face validity’ is used here in its established psychometric sense: the degree to which outputs appear, on inspection, consistent with clinical logic and expert expectation. It does not imply predictive validity, diagnostic accuracy, or concordance beyond what is explicitly reported below.

## 4.1 Retrospective Case-Series Application

The case series was drawn from consecutive clinical records at a single prosthetic-orthotic centre (Ortho Med O&P, Damascus, Syria; January 2023–June 2024). Of 87 consecutive records reviewed, 30 met inclusion criteria (a documented orthotic prescription by a certified orthotist, complete input data, and de-identification); the 57 exclusions were predominantly incomplete records (n = 38), post-surgical presentations (n = 12), or primarily prosthetic cases (n = 7). The final cohort spanned drop foot (n = 6), flexible flatfoot (n = 5), cerebral palsy (n = 5), rigid flatfoot (n = 4), gastrocnemius tightness (n = 4), and poliomyelitis (n = 3), with two overlapping presentations classified by primary indication. Clinical inputs were extracted from the record by the author and independently entered by a research assistant blinded to the original prescription; discrepancies were resolved by consensus reference to the source record.

System-generated suitability scores ranged 76–97% (mean ≈ 89.5%, SD ≈ 5.2%). The lowest score (76%, case P008) occurred in severe spasticity (Ashworth = 4), where the system down-ranked the standard posterior leaf-spring AFO and emitted a caution flag, correctly prioritising patient safety over forced matching. High-match cases (97%) included straightforward poliomyelitis and flaccid drop-foot presentations, and a rigid-flatfoot case in which a negative Jack's test correctly triggered hard exclusion of rigid and semi-rigid insoles in favour of a soft accommodative design. The remaining 27 cases clustered at 82–95%, reflecting mixed indications without hard exclusions.

**Concordance with the original prescription.** Because the source records also document each patient's actual orthotist-issued prescription, we additionally computed agreement between Orthotica's leading recommendation and that prescription. Device-category agreement was 30/30 (100%); exact-subtype agreement was 28/30 (93%), with the two discrepancies reflecting instances where the orthotist's free-text label did not specify the exact subtype distinction captured by the system's structured taxonomy, rather than a substantive disagreement. This concordance should be read cautiously: the rule base was itself shaped, in part, by informal consultation with clinicians from the same regional professional

community as the source clinic, so a high agreement rate partly reflects shared provenance rather than an independent test. It is reported as a descriptive, exploratory statistic, not as diagnostic accuracy or as sensitivity/specificity against a genuine external reference standard.

These retrospective results establish that the rule engine performs as designed and shows exploratory concordance with one clinic's prescribing pattern not that its recommendations would agree with independent prescribers elsewhere or predict real-world clinical outcomes.

# 4.2 Structured Expert Review

Ten practitioners were purposively sampled for professional diversity: 4 certified orthotists, 2 rehabilitation physicians, 2 physical therapists with orthotics training, 1 orthotics researcher/biomechanist, and 1 clinical educator (experience 3–25 years, median 12). Geographic representation was concentrated in one region (Syria n = 4, Jordan n = 2, Lebanon n = 2, Egypt n = 1, United Kingdom n = 1); 8 of 10 were native Arabic speakers, identified through ISPO and national professional networks. Panellists were not blinded to the tool's identity or developer, which may have introduced positive bias, and no comparator system was evaluated in parallel. Each received a technical description, access to three synthetic case scenarios distinct from the retrospective series, and a standardised evaluation instrument (Supplementary Appendix 2) combining 10-point ordinal ratings across five domains, Likert-scale items, open-ended questions, and a demographic questionnaire. Response rate was 100% (10/10) over a two-week window.

Ratings clustered narrowly at the high end of the scale (8–10/10 on every domain; Table 1), with a mean overall-utility rating of 8.9/10 (SD 0.74). This ceiling-effect pattern, combined with the absence of blinding and the geographically narrow, professionally connected sample, means these ratings are better read as a favourable regional face-validity signal than as evidence of broad clinical acceptance. The single-measures intraclass correlation coefficient (ICC) was 0.72 (95% CI 0.51–0.89) a wide interval whose lower bound approaches poor agreement and the average-measures ICC was 0.94; Kendall's coefficient of concordance (W) across the five domains was 0.68 ($p < 0.001$). Given the small, non-representative panel and the absence of a comparator system, these statistics should be treated as exploratory rather than generalisable.

Thematic analysis of open-ended responses converged on three themes: transparency and traceability, appropriately conservative contraindication flagging, and bilingual utility, alongside two explicitly raised methodological concerns that rule weights appeared author-assigned without formal consensus, and that the 30-case retrospective series alone could not establish more than internal consistency. Both concerns are addressed in § 8 and § 9.

These findings establish a preliminary, regionally limited signal of face validity and perceived educational value; they do not establish predictive validity, diagnostic accuracy, or generalisable clinical endorsement.

Table 1
Individual expert ratings across five evaluation domains (n = 10 panellists)

| Panellist | Overall Utility | Transparency of Reasoning | Conservative Safety Logic | Bilingual Interface | Educational Value |
|---|---|---|---|---|---|
| E01 | 9 | 9 | 9 | 9 | 9 |
| E02 | 8 | 9 | 9 | 8 | 9 |
| E03 | 9 | 9 | 9 | 9 | 9 |
| E04 | 10 | 10 | 10 | 9 | 10 |
| E05 | 8 | 8 | 8 | 8 | 8 |
| E06 | 9 | 9 | 9 | 9 | 9 |
| E07 | 9 | 9 | 9 | 8 | 9 |
| E08 | 8 | 9 | 9 | 9 | 8 |
| E09 | 9 | 10 | 9 | 9 | 9 |
| E10 | 10 | 10 | 10 | 10 | 10 |
| Mean | **8.9** | **9.2** | **9.1** | **8.8** | **9.0** |

*Note. E01–E10 are anonymised panellist identifiers. Ratings on 10-point ordinal scales. Single-measures ICC = 0.72 (95% CI 0.51–0.89); average-measures ICC = 0.94 (95% CI 0.87–0.98); Kendall's W = 0.68 ($p < 0.001$), computed across the five domains. All statistics are exploratory given the small, non-representative panel.*

# 5. WEIGHT-SENSITIVITY ANALYSIS

A key limitation of any rule-based system with author-assigned weights is the sensitivity of its outputs to those weights. We conducted a systematic ± 20% perturbation analysis on five representative cases spanning the cohort's clinical range: a straightforward high-match case (poliomyelitis, 97%), a straightforward AFO case (flaccid drop foot, 97%), a conservative-flagging case (severe spasticity, 76%), an intermediate cerebral palsy case (crouch gait, 89%), and a rigid-flatfoot hard-exclusion case (97%). Every non-zero rule weight was perturbed independently by + 20% and − 20%, and the leading recommendation at each layer was recorded; stability was defined as preservation of that leading recommendation. This approach is intentionally simple it does not model correlated perturbations and provides a reproducible lower-bound estimate; Monte Carlo simulation with assumed weight distributions is planned for the Delphi-calibration phase (§ 9).

Across 1,240 individual perturbations (5 cases × 2 directions × 124 non-zero weights), the leading recommendation was preserved in 80.2% (995/1,240) overall; this perturbation-weighted figure is used throughout this paper as the single headline stability statistic, since it counts every perturbation once regardless of how many layers were active for a given case. Instability concentrated in cases with

closely competing alternatives (raw score gap < 5 points between first and second rank), which accounted for 87% of all rank changes despite representing only 23% of perturbation conditions.

Layer-specific stability, averaged per layer across the cases in which that layer was activated (Table 2), varied from 98% (device category, Layer 1) to 71% (AFO subtype, Layer 2), with intermediate results for ankle-joint mechanisms (74%, Layer 3) and insole subtypes (97%, Layer 4, driven by Jack's test acting as a hard-exclusion rule). The heightened sensitivity in Layer 2 reflects genuine clinical overlap: tone-reducing, solid-rigid, and plantarflexion-stop subtypes can all attract positive weight from partially overlapping indicators (moderate spasticity, preserved proximal strength, mild fixed equinus), so small weight changes shift the balance between clinically plausible alternatives. This is an honest reflection of evidence ambiguity in the orthotics literature rather than a design flaw, but it means AFO subtype rankings specifically should not be treated as validated pending formal weight calibration.

High-level recommendations (device category, insole type) are therefore reasonably robust to moderate weight perturbation, while the AFO subtype layer is genuinely sensitive reinforcing the need for Delphi-based calibration before that layer can claim clinical validity. Users are shown, via the inline rationale panel, whenever the score gap to the closest alternative is small (< 5 points), signalling that the ranking warrants heightened caution.

Table 2
Weight-sensitivity analysis by layer and case (n = 5 representative cases)

| Case | Layer 1 (Category) | Layer 2 (AFO Subtype) | Layer 3 (Joint Mechanism) | Layer 4 (Insole) |
|---|---|---|---|---|
| P006 (Poliomyelitis, 97%) | 100% | 85% | N/A | 95% |
| P030 (Drop foot, 97%) | 100% | 72% | N/A | N/A |
| P008 (Spasticity, 76%) | 92% | 58% | 65% | N/A |
| P017 (CP, crouch, 89%) | 96% | 68% | 82% | N/A |
| P023 (Rigid flatfoot, 97%) | 100% | N/A | N/A | 98% |
| Layer mean | **98%** | **71%** | **74%** | **97%** |

*Note. Stability = percentage of ± 20% weight perturbations that preserved the original leading recommendation. N/A = layer not activated for that case. Layer means are the simple arithmetic mean of the case values shown (N/A cells excluded). The single headline overall-stability figure reported in the text (80.2%) is computed across all 1,240 individual perturbations directly, not as an average of this table's per-case values, and is the only overall figure used in this paper.*

# 6. DISCUSSION

Orthotica's principal contribution is a structured, transparent, freely available way of organising the many factors bearing on foot-and-ankle orthosis selection, together with an honest, quantified account of where its current weights are and are not stable. The retrospective series shows the rule engine performing as designed, and the concordance analysis provides an exploratory, if provenance-limited, signal that its recommendations track a working clinician's actual prescribing pattern at the category level. The sensitivity analysis adds a dimension rarely reported in CDSS prototypes: an explicit, reproducible quantification of how much a recommendation depends on the specific weights chosen, which is precisely the information needed to prioritise where formal consensus calibration matters most (Layer 2).

The choice of a rule-based architecture positions Orthotica within the interpretable-CDSS tradition, in deliberate contrast to black-box approaches increasingly common in medical artificial intelligence [8]. For a tool intended to scaffold, not replace, the judgment of a certified orthotist and physician, interpretability is a safety requirement, not a stylistic preference. Sutton and colleagues identify successful CDSS as depending on embedding within a defined clinical workflow, delivery at the point of decision-making, and actionable, non-intrusive presentation [8]; Orthotica satisfies the latter two through offline Progressive Web App delivery and the rationale panel, while workflow embedding remains to be established through prospective studies. Within the orthotics literature, recent systematic reviews converge on the same structural observation that the cumulative evidence is heterogeneous and difficult to translate at the point of individual prescription [5–7]. Orthotica does not resolve this heterogeneity; it consolidates its consequences into a single inspectable, contestable artefact, which is a precondition for the empirical resolution the field needs.

Four limitations of proportionally greater importance than the rest warrant emphasis here (formal limitations follow in § 8). First, 'suitability match' percentages are ordinal heuristics of rule agreement, not calibrated probabilities of clinical success a 97% score means the presentation matched more positive-indication rules than any alternative, nothing more. Second, weights were assigned by the author with informal clinician input rather than a Delphi process, and § 5 confirms that AFO subtype rankings are the most sensitive to this. Third, the embedded clinical measures (Ashworth, FPI-6, navicular drop, Silfverskiöld) carry their own documented reliability limits [10–14], so measurement noise upstream propagates into the recommendation regardless of rule-base quality. Fourth, the retrospective series and concordance figures come from thirty patients at a single centre selected by record completeness, so both figures describe internal and provenance-limited consistency, not generalisable accuracy or agreement with independent prescribers elsewhere.

Given these limits, we regard Orthotica's present, appropriate uses as organisational and educational a structured checklist that helps ensure no relevant clinical dimension is overlooked, and a teaching aid whose exposed rule logic lets trainees see how findings translate into a candidate device rather than a tool ready to independently inform prescribing decisions ahead of formal calibration and prospective, blinded validation.

# 7. CONCLUSION

We have described Orthotica, a transparent, rule-based, bilingual web application designed to structure and document the reasoning behind foot-and-ankle orthosis selection. A preliminary retrospective assessment on 30 clinical cases showed internally consistent, conservatively flagged recommendations and exploratory concordance with one clinic's prescribing pattern (100% category-level, 93% subtype-level), a finding limited by shared provenance between the rule base and that clinic. A weight-sensitivity analysis showed device-category recommendations are robust to moderate perturbation (98% stability) while AFO subtype rankings are markedly more sensitive (71%), reinforcing the priority for formal Delphi-based calibration. A ten-practitioner expert review, drawn largely from one region and unblinded to the tool's identity, yielded high but provenance-limited face-validity ratings (mean 8.9/10; ICC = 0.72, 95% CI 0.51–0.89). Orthotica is offered as an engineering, methodological, and educational prototype whose rules are literature-grounded and practitioner-informed but not yet derived from formal consensus or tested against prospective, blinded, independent outcomes. Its design, data, and complete rule base (Supplementary Appendix 1) are published in full to invite the collaborative, multi-institutional refinement needed to move this transparent heuristic toward a validated clinical aid.

# 8. LIMITATIONS

This work has several limitations that should be considered when interpreting its findings.

**Sample size and representativeness.** The retrospective case series comprises 30 cases across six diagnostic categories, drawn from consecutive records at a single centre and selected by record completeness. This is adequate for qualitative internal-consistency observation but not for quantitative conclusions about generalisability, and the same constraint applies to the concordance figures in § 4.1.

**Reference standard.** No independent, blinded reference standard was used. Concordance with the source clinic's own orthotist is provenance-limited, given that the rule base's own development history involved informal consultation with clinicians from the same regional professional community (§ 4.1); it is not a substitute for validation against outcomes at independent centres.

**Weight provenance.** Rule weights are author-assigned and informed by informal practitioner consultation; no structured rating instrument or inter-rater agreement statistic was computed at development time. § 5 quantifies, but does not resolve, this limitation.

**Expert-review methodology.** The ten-member panel was not blinded to the tool's identity or developer, no comparator system was evaluated, and the sample was geographically narrow and professionally connected (predominantly the Levant region). Ratings clustered near the ceiling of the scale, and the ICC's lower confidence bound (0.51) approaches poor agreement. These ratings should not be generalised beyond the sampled population.

**Embedded measurement reliability.** The clinical inputs the system accepts Ashworth [10], FPI-6 [11, 12], navicular drop [13], Silfverskiöld [14], carry documented reliability constraints of their own, which propagate into the system's recommendations and are visible to the user only via inline interpretive guides.

**Cultural representativeness of the knowledge base.** Rule content draws predominantly on English-language consensus and systematic-review literature; Arabic-language clinical-evidence sources were not separately surveyed for this first iteration.

**No outcome data.** No patient-outcome data have been collected. A 97% suitability score does not imply a 97% probability of favourable clinical outcome; this distinction is surfaced in the user interface and reiterated throughout this manuscript.

**Regulatory status.** While Orthotica is positioned as a low-risk decision-support tool (IMDRF SaMD Category A) [15], it has not undergone formal regulatory review, clinical-risk assessment, or post-market surveillance. The clinical responsibility model that the tool scaffolds rather than replaces clinical judgment is a design intention, not a legally tested framework.

# 9. FUTURE WORK: DELPHI CALIBRATION AND PROSPECTIVE VALIDATION

To make the proposed consensus process concrete and time-bound, we outline a protocol that will be reported in full, following CREDES guidance [16], in a subsequent paper.

**Panel and rounds.** Fifteen to twenty internationally recognised experts (≥ 50% certified orthotists; remainder rehabilitation physicians, biomechanists, physical therapists), purposively sampled for geographic diversity including at least three Arabic-speaking jurisdictions beyond the Levant. Round 1 uses a structured online questionnaire rating each rule's clinical importance on a 9-point scale and inviting additions or splits. A pre-defined consensus threshold of ≥ 70% agreement is applied after each round, with Krippendorff's α (ordinal) used for inter-rater agreement rather than simple percentage agreement; rules below threshold are re-presented with anonymised aggregated feedback in Round 2. Round 3 confirms final agreement statistics; the resulting weight schema is published as a structured supplementary file, and any statement with persistent disagreement (> 30% dissent) is documented as an unresolved uncertainty rather than silently resolved by majority vote.

**Prospective validation.** Following calibration, a prospective, blinded concordance study at ≥ 2 independent centres (target n ≥ 200 new prescriptions) will compare Orthotica's leading recommendation against a board-certified orthotist's independent prescription, blinded to the tool's output, using Cohen's κ as the primary outcome, stratified by diagnostic category as a secondary outcome. Success will be defined a priori as $\kappa \geq 0.60$ (substantial agreement) for device-category recommendations; falling short of this threshold will be reported and discussed rather than reframed post hoc.

Funding, ethics approval, and resourcing for this multi-centre programme (Table 3) have not yet been secured; the timeline below is a proposal, not a committed plan, and its Phase 6 prospective study requires ethical approval and formal institutional agreements at each participating centre before enrolment begins.

Table 3
Proposed timeline for Delphi calibration and prospective validation

| Phase | Activity | Duration |
|---|---|---|
| Phase 1 | Panel recruitment & instrument piloting | 3 months |
| Phase 2 | Delphi Round 1 (open elicitation) | 6 weeks |
| Phase 3 | Delphi Round 2 (structured re-rating) | 4 weeks |
| Phase 4 | Delphi Round 3 (consensus confirmation) | 4 weeks |
| Phase 5 | Weight schema publication & system update | 2 months |
| Phase 6 | Prospective concordance study (n ≥ 200) | 12 months |

*Note. Timeline is approximate and contingent on funding and centre recruitment. Phase 6 requires ethical approval and institutional agreements at participating centres.*

## Declarations

A statement of ethics approval: This paper describes software engineering, knowledge-based design, retrospective analysis of de-identified data, and structured expert feedback. No prospective human participants, interventional patient data, or identifiable patient records were involved at any stage; All retrospective cases were de-identified prior to analysis.

# Conflicts of interest:

The author is the sole designer and developer of the software described and has a direct interest in its description; no commercial entity was involved in its development, and the application is distributed free of charge.

### Ethics

This paper describes software engineering, knowledge-based design, retrospective analysis of de-identified data, and structured expert feedback. No prospective human participants, interventional patient data, or identifiable patient records were involved at any stage; no ethical approval was therefore required under the institutional guidelines applicable at the author's centre. All retrospective cases were de-identified prior to analysis.

# Funding:

This work received no external funding and was undertaken independently by the author.


# Author contributions:

Mohamad Firas Wahbeh: conceptualisation, methodology, investigation, funding acquisition, project administration, supervision, writing – original draft.

# Data availability:

The complete rule base and weight schema are provided as Supplementary Appendix 1 (Microsoft Excel format). The application source code is available upon reasonable request to the corresponding author. The retrospective case data are not publicly available due to institutional data-sharing restrictions, but de-identified summary statistics are reported in § 4.1 and Supplementary Appendix 1 (Scoring Examples sheet).

## References


1. Condie DN, Meadows CB (eds) (1995) Report of a Consensus Conference on the Lower Limb Orthotic Management of Cerebral Palsy. Copenhagen: International Society for Prosthetics and Orthotics
2. International Society for Prosthetics and Orthotics. ISPO Consensus Conference on Poliomyelitis: consensus statements. Prosthet Orthot Int (2001) ;25:171–80
3. Condie DN (2008) International Organization for Standardization terminology. Atlas of Orthoses and Assistive Devices, 4th edn. Mosby Elsevier, Philadelphia, pp 3–7
4. Morris C (2002) A review of the efficacy of lower-limb orthoses used for cerebral palsy. Dev Med Child Neurol 44:205–211
5. Figueiredo EM, Ferreira GB, Maia Moreira RC, Kirkwood RN, Fetters L (2008) Efficacy of ankle-foot orthoses on gait of children with cerebral palsy: systematic review of literature. Pediatr Phys Ther 20:207–223
6. Ridgewell E, Dobson F, Bach T, Baker R (2010) A systematic review to determine best practice reporting guidelines for AFO interventions in studies involving children with cerebral palsy. Prosthet Orthot Int 34:129–145
7. Miccinilli S, Santacaterina F, Della Rocca R, Sterzi S, Bressi F, Bravi M (2024) Efficacy of lower limb orthoses in the rehabilitation of children affected by cerebral palsy: a systematic review. Child (Basel) 11:212

8. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI (2020) An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med 3:17
9. Palisano R, Rosenbaum P, Walter S, Russell D, Wood E, Galuppi B (1997) Development and reliability of a system to classify gross motor function in children with cerebral palsy. Dev Med Child Neurol 39:214–223
10. Bohannon RW, Smith MB (1987) Interrater reliability of a modified Ashworth scale of muscle spasticity. Phys Ther 67:206–207
11. Redmond AC, Crosbie J, Ouvrier RA (2006) Development and validation of a novel rating system for scoring standing foot posture: the Foot Posture Index. Clin Biomech (Bristol Avon) 21:89–98
12. Keenan AM, Redmond AC, Horton M, Conaghan PG, Tennant A (2007) The Foot Posture Index: Rasch analysis of a novel, foot-specific outcome measure. Arch Phys Med Rehabil 88:88–93
13. Mueller MJ, Host JV, Norton BJ (1993) Navicular drop as a composite measure of excessive pronation. J Am Podiatr Med Assoc 83:198–202
14. Martinez L, Lalevée M, Beldame J, L'Hermette M, Brunel H, Dujardin F et al (2023) Reliability of a new computerized equinometer based on Silfverskiöld test to measure gastrocnemius tightness. PLoS ONE 18:e0284279
15. International Medical Device Regulators Forum Software as a Medical Device (SaMD): key definitions. IMDRF/SaMD WG/N10FINAL:2013 [Internet]. 2013 [cited 2026 Aug 9]. Available from: https://www.imdrf.org/documents/software-medical-device-samd-key-definitions
16. Jünger S, Payne SA, Brine J, Radbruch L, Brearley SG (2017) Guidance on Conducting and REporting DElphi Studies (CREDES) in palliative care: recommendations based on a methodological systematic review. Palliat Med 31:684–706

# Figures




**Figure 1**

*Figure 1. Application entry screen showing the design philosophy, workflow navigation, and live suitability engine.*

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryAppendix1.xlsx
- CaseSeriesDataset.xlsx
- SupplementaryAppendix2.docx