

Table of Contents

Evidenzgewichtete Softwarearchitektur für integrative Naturstoff- und Biomarkerforschung

Entwurfsprinzipien, Bewertungssemantik, regulatorische Einordnung und Prüfprotokoll

Thomas Schüller · Syntagma Forschung, Wien, Österreich forschung@tschueller.com · ORCID 0009-0003-9799-6747 Unabhängiger Forscher. Privat finanziert, ohne institutionelle Affiliation, ohne Interessenkonflikte.

Version 0.3 — Preprint — 2026-09-04 · Lizenz CC BY 4.0

Kurzfassung

Die Arbeit beschreibt den Entwurf einer Forschungsplattform (BioLogic × FBS), die Naturstoffforschung und individuelle Biomarkerdaten in einem evidenzgewichteten Wissensmodell zusammenführt, und benennt die Bedingungen, unter denen dieser Entwurf geprüft oder widerlegt werden kann. Das System übernimmt strukturierte Datensätze aus offenen wissenschaftlichen Quellen (PubMed, ChEMBL, UniProt, LOTUS, ClinicalTrials.gov, FDA FAERS, Reactome, Open Targets und weitere), normalisiert chemische Identität über InChI/SMILES, codiert Biomarkerwerte über LOINC und verknüpft Entitäten in einem Eigenschaftsgraphen mit vier primären Knotentypen und neun Beziehungstypen. Jede Beziehung trägt einen Evidenzsatz: klassenabgeleitetes Grundgewicht, multiplikative Qualitätsmodifikatoren, abgeleitete Unsicherheitsstufe, Provenienz.

Version 0.3 nimmt drei Festlegungen aus v0.1 zurück. Erstens wird das Bewertungsverfahren nicht mehr als GRADE-Anpassung geführt: GRADE bezeichnet sich selbst als nicht-quantitatives System und bewertet einen Evidenzkörper je Endpunkt, während dieses System ein Zahlengewicht je Datensatz erzeugt. Das Verfahren wird daher als GRADE-informiert und ausdrücklich nicht GRADE-konform neu gefasst, mit aufgezählten Unterschieden. Zweitens ergänzt die Arbeit ein präregistrierbares Kalibrierungsprotokoll mit Falsifikationskriterien; damit wird die größte Einschränkung von v0.1 aus einem Eingeständnis in einen prüfbaren Plan überführt. Drittens ergänzt sie eine regulatorische Einordnung zu Softwarequalifikation nach MDR/IVDR gemäß MDCG 2019-11 Rev. 1, zum Anwendungsstand der Verordnung (EU) 2024/1689 (KI-Verordnung) nach der Änderung von 2026, zu Art. 9 DSGVO und zum Zeitplan des Europäischen Gesundheitsdatenraums — jeweils mit dem Auslöser, der die Einordnung kippen würde.

Das System bleibt ein Forschungsinstrument. Es erzeugt Assoziationen mit Unsicherheitskennzeichnung, keine Diagnosen, keine Therapieempfehlungen, keine Kausalaussagen. Zwölf Einschränkungen sind dokumentiert. Eine empirische Kalibrierung liegt

nicht vor; alle Werte der laufenden Version sind als Ordnungsheuristik zu lesen, nicht als Wahrscheinlichkeiten.

Schlagwörter: Evidenzgewichtung, Naturstoffe, Biomarker, Wissensgraph, Provenienz, Reproduzierbarkeit, Softwarequalifikation, Forschungssoftware

1. Einleitung

1.1 Problemstellung

Integrative Naturstoffforschung verlangt den gleichzeitigen Umgang mit ungleichartigen Datensätzen: chemische Strukturen, Zielproteine, Signalwege, präklinische und klinische Studien, Nebenwirkungsmeldungen, individuelle Biomarkerwerte, rechtlicher Kontext. Offene Datenbanken decken jede dieser Domänen für sich gut ab. Was fehlt, sind nicht Daten, sondern eine begründbare Art, Datensätze ungleichen Evidenzrangs in einer abfragbaren Struktur zu halten, ohne sie stillschweigend einzuebnen.

Zwei Fehlermodi begründen den Entwurf. Der erste ist *Einebnung*: ein In-vitro-IC50 und eine randomisierte kontrollierte Studie stehen als Nachbarzeilen derselben Tabelle und werden als gleichwertig gelesen. Der zweite ist *Übertragung*: eine an einem standardisierten Extrakt gezeigte Eigenschaft wird auf die Reinsubstanz übertragen, oder eine im Nagermodell gezeigte Wirkung wird als Wirkung beim Menschen gesprochen. Beide Fehlermodi sind strukturell, nicht zufällig — sie folgen aus Datenmodellen, die die Behauptung speichern und die Bedingungen ihrer Erzeugung fallen lassen.

1.2 Beitrag

1. Ein Domänenmodell über Naturstoffchemie, Zielstrukturen, Signalwege und individuelle Biomarkerwerte, mit dem Substanztyp als erstrangigem, nicht übertragbarem Attribut (Extrakt / Reinsubstanz / Derivat / theoretisches Konstrukt).
2. Ein Evidenzgewichtungsverfahren mit deterministischen, prüfbaren, versionierten Werten je Beziehung, zusammen mit der ausdrücklichen Angabe, was die Zahlen *nicht* bedeuten (§4.1, §4.7).
3. Ein vierstufiges Qualitätstor mit Quarantäne statt Löschung, sodass abgewiesene Datensätze einsehbar bleiben.
4. Ein Kalibrierungsprotokoll mit vorab festgelegten Erfolgs- und Misserfolgskriterien (§5) — der Teil der Arbeit, der den Bewertungsentwurf tatsächlich widerlegen kann.
5. Eine regulatorische Einordnung mit benannten Auslösebedingungen je Einordnung (§7) anstelle des reinen Haftungsausschlusses aus v0.1.
6. Ein Prüf- und Betaprotokoll mit Datenminimierung und Abbruchregeln (§8).
7. Zwölf dokumentierte Einschränkungen (§9).

1.3 Änderungen gegenüber v0.1

Bereich	v0.1 (2026-05-14)	v0.3 (2026-09-04)
Rahmung der Bewertung	„GRADE-adapted quantitative framework“	GRADE-informiert, ausdrücklich nicht konform;

Bereich	v0.1 (2026-05-14)	v0.3 (2026-09-04)
		Unterschiede aufgezählt (§4.1)
Bezugsobjekt	nicht benannt	benannt: Datensatzebene, nicht Evidenzkörper (§4.2)
Kalibrierung	Einschränkung L1, kein Plan	präregistrierbares Protokoll mit Falsifikationskriterien (§5)
Regulatorik	ein Satz Haftungsausschluss	Qualifikationsanalyse mit Auslösetabelle (§7)
Prüfung	51 Unit-Tests	Tests plus externes Prüf-/Betaprotokoll (§8)
Einschränkungen	6	12
Quellen	12	28, geprüft am 2026-09-04
Sprachen	nur EN	EN + DE gleichrangig

1.4 Geltungsbereich und ethische Grenze

Das System dient Forschung, Hypothesenbildung und Wissenserschließung. Es diagnostiziert nicht, empfiehlt keine Behandlung und behauptet keine Kausalität. Jede Ausgabe trägt eine Unsicherheitskennzeichnung. Militärische, waffentechnische und wehrtechnische Anwendungen sind im gesamten Vorhaben ausgeschlossen, ausnahmslos und unabhängig von Finanzierung.

2. Verwandte Arbeiten und Abgrenzung

Die Evidenzhierarchie ist von GRADE [1] und vom Cochrane-Bias-Rahmenwerk [2, 3] informiert. Chemische Identität folgt InChI [4] und SMILES [5]; Laborbeobachtungen folgen LOINC [6]; Provenienz folgt W3C PROV-DM [7]; der Datenumgang folgt den FAIR-Prinzipien [8].

Im Naturstoffbereich liefert LOTUS [9] offene Substanz-Organismus-Verknüpfungen, ChEMBL [10, 11] kuratierte Bioaktivitätsdaten, Reactome [12] Signalwegwissen, Open Targets [13] aggregierte Ziel-Krankheits-Assoziationen mit eigenem Bewertungsmodell und UniProt [14] Proteinreferenzdaten. Pharmakovigilanzsignale stammen aus FDA FAERS, dessen bekannte Meldeverzerrungen als Modifikator behandelt werden, nicht als Wirknachweis [15].

Abgrenzung. Das Unterscheidungsmerkmal ist nicht die Abdeckung — jede genutzte Quelle ist öffentlich —, sondern dass eine einheitliche Gewichtungs- und Unsicherheitssemantik über bevölkerungsbezogene und individuelle Datensätze hinweg angewandt und in angreifbarer Form aufgeschrieben wird.

Eigenschaft	ChEMBL	Open Targets	Reactome	LOTUS	FHIR/LIS	dieses System
Kuratierte Bioaktivität	ja	über ChEMBL	nein	teilweise	nein	importiert
Ziel-Krankheits-Bewertung	nein	ja (eigenes Modell)	nein	nein	nein	ja (eigenes Modell)
Naturstoff-Provenienz	Kennzeichen	nein	nein	ja	nein	ja
Individuelle Biomarkerwerte	nein	nein	nein	nein	ja	ja
Einheitliche Unsicherheit je Datensatz	nein	teilweise	nein	nein	nein	ja
Substanztyp-Übertragung gesperrt	nein	nein	nein	teilweise	entfällt	ja
Werte empirisch kalibriert	entfällt	teilweise	entfällt	entfällt	entfällt	nein (\$5)

Die letzte Zeile ist die ehrliche. Open Targets gewichtet Evidenz nach Datentyp in einem dokumentierten Modell [13]; dieses System tut Vergleichbares in einer anderen Domäne, hat aber noch nicht gezeigt, dass seine Gewichte irgendetwas nachbilden.

3. Systemarchitektur

3.1 Quellen

55 Quellen aus zehn Domänen sind im beiliegenden Quellenregister dokumentiert, jeweils mit Zugriffsweg, Lizenz, Aktualisierungsfrequenz und Qualitätsstufe A (begutachtet oder Standard), B (kuratiert, etabliert), C (Community oder Aggregator), D (veraltet oder eingeschränkt). Das Register hält je Quelle fest, ob sie in der laufenden Prüfrunde direkt verifiziert oder aus v0.1 übernommen wurde; übernommene Einträge sind gekennzeichnet und gelten nicht als verifiziert.

3.2 Domänenmodell

24 Entitätstypen in zwei Gruppen.

Naturstoff-/Discovery-Seite (FBS, 12): Compound (InChIKey-indiziert), Organism (NCBI-Taxonomie), Target (UniProt-indiziert), Pathway, Assay, Study, Observation, AdverseEvent, Patent, Quarantine, IngestionLog sowie Evidenzmetadaten auf jeder Beziehung.

Individuelle Seite (BioLogic, 12): User, UserProfile, LabReport, BiomarkerReading (LOINC-codiert), BiomarkerTimeseries, Hypothesis, HypothesisEvidence, LabPartner, Sample, GeneratedReport, DataProvenance, AuditLog.

Der Substanztyp ist Attribut der Compound-Entität und jeder von ihr ausgehenden Beziehung. Regel C3 (§4.6) verbietet die Weitergabe extraktbasierter Evidenz an Reinsubstanzen; die Regel greift beim Schreiben, nicht erst bei der Anzeige.

3.3 Wissensgraph

Vier primäre Knotentypen (Compound, Target, Pathway, Biomarker/Analyt) und ergänzende Typen (System, Nutrient, Gene, Mechanism, Publication). Neun Beziehungstypen. Jede Kante trägt: `evidence_class`, `base_score`, `applied_modifiers[]`, `final_score`, `uncertainty`, `source_reference`, `config_version`, `provenance_timestamp`. Eine Kante ohne vollständigen Evidenzsatz kann nicht geschrieben werden; die Aufnahmeschicht weist sie in die Quarantäne.

3.4 Technik und Reproduzierbarkeit

FastAPI (Python 3.12+), PostgreSQL 16, Neo4j 5.x, Celery/Redis, React/TypeScript, RDKit, MeiliSearch, containerisiert über Docker Compose. Die Bewertungskonfiguration ist eine versionierte Datei; die Version wird auf jeden berechneten Wert gestempelt, sodass jeder Wert nachgerechnet und gegen die Konfiguration seiner Entstehung geprüft werden kann.

3.5 Modulgrenzen

Die Modulgrenzen verlaufen entlang der regulatorischen Bruchlinie aus §7, nach der modulweisen Bewertungslogik von MDCG 2019-11 Rev. 1 [16]:

Modul	Funktion	Regulatorischer Charakter (§7)
M1 Aufnahme & Qualitätstor	Abruf, Prüfung, Quarantäne	Infrastruktur
M2 Wissensgraph	Speicherung, Traversierung	Infrastruktur
M3 Evidenzgewichtung	Bewertung, Unsicherheit, Konflikte	Forschungsmethode
M4 Laborimport	Befundeinlesen, LOINC-Codierung	personenbezogene Daten (Art. 9 DSGVO)
M5 Biomarker-Kontext	Wert gegen Referenzbereich anzeigen	Grenzmodul
M6 Hypothesenwerkbank	nutzerseitige Annotation und	Forschungsmethode

Modul	Funktion	Regulatorischer Charakter (§7)
Verknüpfung		
M7 Export & Rechte	Art. 15/17/20, Prüfprotokoll	personenbezogene Daten

M5 entscheidet über den regulatorischen Charakter des gesamten Systems; deshalb ist es isoliert und seine Zweckbestimmung getrennt festgeschrieben (§7.2).

Abb. 1 — Systemarchitektur und Modulgrenzen / System architecture and module boundaries

BioLogic x FBS v0.2 — FORSCH-PROV0427 — CC BY 4.0

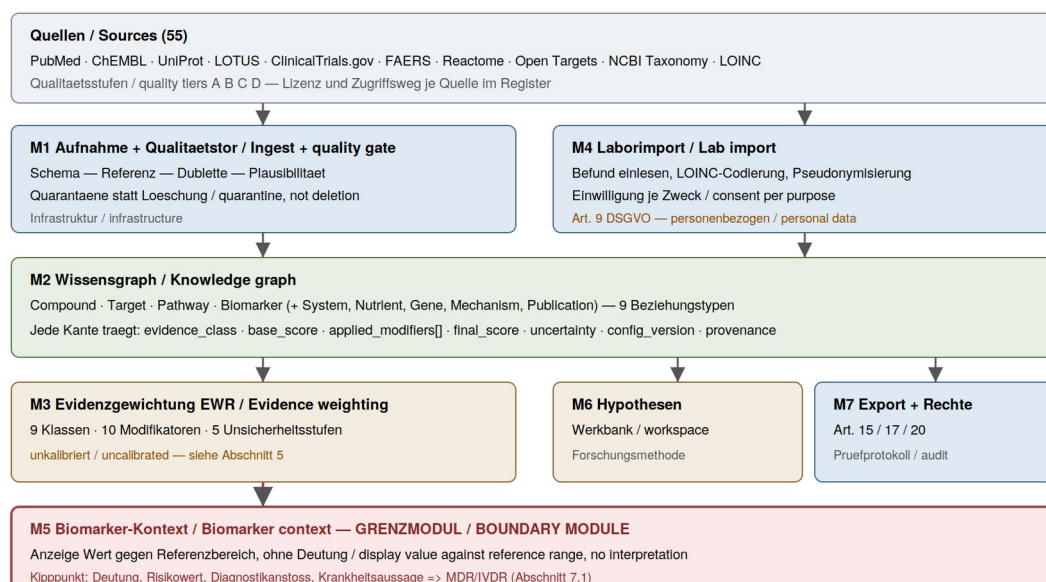


Abbildung 1 — Systemarchitektur mit Modulgrenzen (M1–M7). M5 ist das Grenzmodul aus §7.1.

4. Evidenzgewichtung

4.1 Verhältnis zu GRADE — Korrektur gegenüber v0.1

Version 0.1 bezeichnete das Verfahren als „GRADE-adapted quantitative framework“. Diese Bezeichnung wird zurückgenommen. Das GRADE-Handbuch stellt fest, dass GRADE kein quantitatives System zur Bewertung der Evidenzqualität ist und dass die Auf- und Abwertungsfaktoren Kontinua abbilden, deren Bewertung Urteilsbildung erfordert [1, 17]. Daraus folgen drei Unterschiede, die hier im Haupttext und nicht in einer Fußnote stehen:

1. **Bezugsobjekt.** GRADE bewertet die Sicherheit eines Evidenzkörpers *für einen bestimmten Endpunkt*. Dieses System hängt ein Gewicht an einen *einzelnen Datensatz beziehungsweise eine einzelne Beziehung*. Das sind verschiedene Gegenstände; ein hoher Wert hier sagt nichts über die Sicherheit eines Evidenzkörpers.

2. **Rechenweise.** Abwertung in GRADE ist eine strukturierte Urteilsbildung, kein Produkt von Koeffizienten. Die hier verwendete multiplikative Verknüpfung (§4.4) ist eine technische Vereinfachung mit den in §4.7 aufgeführten Eigenschaften und Mängeln.
3. **Empfehlungsstärke.** GRADE trennt Evidenzsicherheit von Empfehlungsstärke. Dieses System spricht keine Empfehlungen aus; die zweite Hälfte von GRADE hat daher keine Entsprechung und darf nicht hineingelesen werden.

Das Verfahren heißt deshalb **EWR (Evidenzgewichtung auf Datensatzebene)** und wird als GRADE-informiert geführt. Wo GRADE-Begriffe in der Oberfläche erscheinen, erscheinen sie als Zuordnungshinweis und sind als solcher gekennzeichnet.

4.2 Bezugsobjekt

Bewertet wird das Tripel (Quelldatensatz, behauptete Beziehung, Extraktionsvorgang). Die Verdichtung auf Aussageebene — „wie steht die Evidenz für Substanz X an Ziel Y“ — ist ein eigener Vorgang (§4.6, Lückenanalyse) und ist bewusst nicht der Mittelwert der Kantenwerte, weil Kanten keine unabhängigen Stichproben sind: mehrere Arbeiten eines Labors oder eine Arbeit in drei Datenbanken würden den Mittelwert sonst aufblähen.

4.3 Evidenzklassen

Rang	Evidenzklasse	Grundgewicht	GRADE-Hinweis (unverbindlich)
1	Systematische Übersicht / Metaanalyse	1,00	hoch
2	Randomisierte kontrollierte Studie	0,90	hoch
3	Kontrollierte klinische Studie	0,80	mäßig
4	Beobachtungsstudie	0,70	niedrig
5	Präklinisch in vivo	0,50	sehr niedrig
6	Präklinisch in vitro	0,35	sehr niedrig
7	In silico	0,15	entfällt
8	Theoretisch	0,05	entfällt
9	Traditionelle Verwendung, unvalidiert	0,02	entfällt

Die Zahlenabstände sind ein Ordnungsmittel. Sie kodieren, dass eine Metaanalyse über einem In-vitro-Befund steht; sie behaupten nicht, eine Metaanalyse sei 2,86-mal „besser“ als ein In-vitro-Befund. §5 prüft, ob die Ordnung dem Urteil geschulter Bewerter standhält.

4.4 Verknüpfung

Modifikator	Faktor	Begründung
Zurückgezogen	0,00	absorbierender Zustand; Datensatz bleibt, Gewicht null
Bias-Risiko hoch	0,50	RoB-2-/ROBINS-I- Domänenurteil [2, 3]
Bias-Risiko unklar	0,75	unzureichende Berichterstattung
Kleine Stichprobe	0,60	unterhalb klassenspezifischer Schwelle laut Konfiguration
Keine Verblindung	0,70	wo Verblindung möglich wäre
Einzellabor	0,70	kein unabhängiger Standort
Preprint	0,50	Begutachtung nicht abgeschlossen
Interessenkonflikt	0,75	Geldgeber- oder Autoreninteresse erklärt oder erkannt
Standardisiertes Protokoll	1,10	benanntes, veröffentlichtes Protokoll
Unabhängige Replikation	1,30	andere Gruppe, vergleichbarer Aufbau

4.5 Unsicherheit

Die Unsicherheit wird abgeleitet, nie von Hand gesetzt:

- **keine** — Rang ≤ 2 und Endwert $\geq 0,7$ und repliziert und kein Konflikt
- **niedrig** — Rang ≤ 4 und Endwert $\geq 0,5$ und kein Konflikt
- **mäßig** — Rang ≤ 6 und Endwert $\geq 0,3$
- **hoch** — Rang ≤ 7 oder Konflikt vorhanden
- **kritisch** — Rang > 7 oder zurückgezogen oder keine belastbare Evidenz

In der Oberfläche erscheint zuerst die Kennzeichnung, nicht die Zahl. Die Zahl ist auf Abruf mit vollständiger Herleitung verfügbar.

4.6 Konfliktbehandlung

C1 Höherrangige Evidenz geht vor; der niederrangige Datensatz bleibt erhalten und sichtbar. C2 Widersprüche gleichen Rangs werden gekennzeichnet, beide Datensätze bleiben, manuelle

Prüfung wird angefordert. Kein automatischer Sieger. C3 Extraktevidenz überträgt sich nicht auf Reinsubstanzen; die Rückübertragung ist ebenso gesperrt. C4 Tierdaten sind keine Humanevidenz; die Spezies ist Teil des Kantenschlüssels. C5 In-silico-Vorhersagen sind ausschließlich Hypothesengeber und können eine Verdichtung nie über die Stufe mäßig heben.

Die Lückenanalyse kehrt dieselben Regeln um: liegt der höchste verfügbare Rang für eine Aussage bei 6 oder schlechter, wird die Aussage als Evidenzlücke berichtet und nicht als schwacher Befund — das ist eine andere Aussage.

4.7 Formale Eigenschaften und bekannte Mängel

Diese Angaben gehören zum Beitrag, weil sie begrenzen, was §5 überhaupt zeigen kann.

- **Monotonie innerhalb einer Klasse gilt.** Ein zusätzlicher Abwertungsfaktor erhöht den Wert nie.
 - **Klassenübergreifende Monotonie gilt nicht.** Eine RCT mit drei Abwertungen ($0,90 \times 0,5 \times 0,6 \times 0,7 = 0,189$) fällt unter einen In-vitro-Befund mit Aufwertung ($0,35 \times 1,1 = 0,385$). Ob diese Umkehrung erwünscht ist, ist eine empirische Frage und der erste Prüfpunkt in §5.4.
 - **Die multiplikative Verknüpfung unterstellt Unabhängigkeit der Modifikatoren.** Kleine Stichprobe und fehlende Verblindung sind praktisch nicht unabhängig; das Produkt bestraft daher doppelt. Ein log-additives Abschlagsmodell mit gedeckelter Gesamtabwertung ist die in §5.5 vorab festgelegte Alternative.
 - **Aufwertungssättigung.** Zwei Aufwertungen ($1,10 \times 1,30 = 1,43$) sind wegen der Deckelung nur für Grundgewichte $\leq 0,70$ wirksam. Aufwertungen sind damit strukturell schwächer als Abwertungen. Das ist beabsichtigt, aber ungeprüft.
 - **Nichtidentifizierbarkeit.** Mehrere Modifikatorkombinationen führen zum selben Endwert. Prüfgegenstand ist das gespeicherte Feld `applied_modifiers[]`, nicht der Skalar.
 - **Keine Unsicherheitsfortpflanzung entlang von Pfaden.** Eine Zweischritt-Traversierung (Substanz → Ziel → Signalweg) meldet derzeit den kleinsten Kantenwert; konservativ, aber grob. Pfadbezogene Unsicherheit ist auf v2 verschoben und als L8 geführt.
-

Abb. 2 — Evidenzgewichtung EWR / Evidence weighting EWR

GRADE-informiert, nicht GRADE-konform / GRADE-informed, not GRADE-conformant

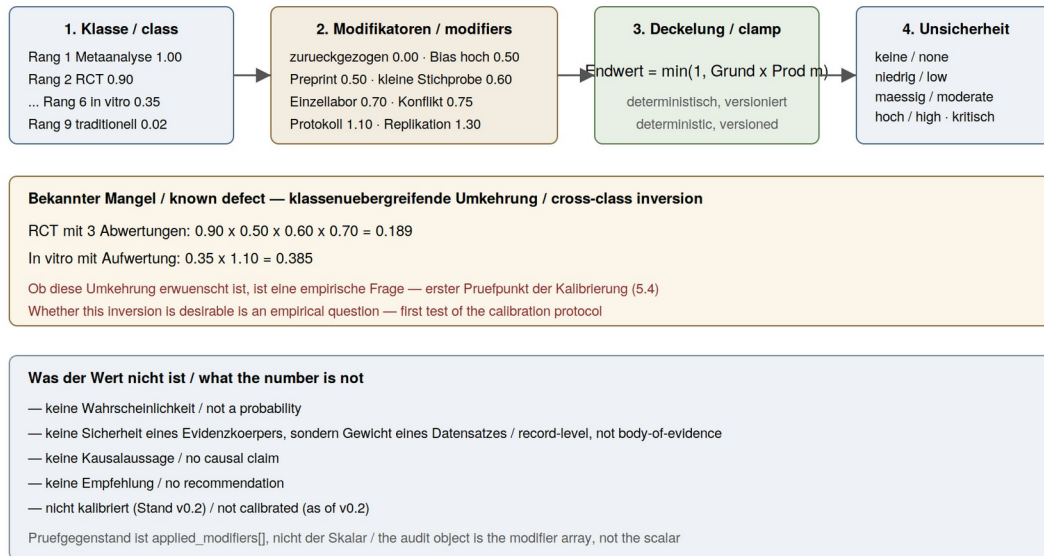


Abbildung 2 — Evidenzgewichtung EWR mit der klassenuebergreifenden Umkehrung aus §4.7.

5. Kalibrierungsprotokoll

Dieser Abschnitt ist in v0.3 neu und ist der Teil, der den Entwurf widerlegen kann. Er ist so geschrieben, dass er vor der Datenerhebung präregistriert werden kann.

5.1 Gegenstand

Geprüft wird, ob EWR-Gewichte Evidenz so ordnen wie geschulte menschliche Bewerter, und ob die Modifikatorstruktur Information über den Klassenrang hinaus trägt.

5.2 Referenzmenge

Eine geschichtete Menge von 300 Beziehungsdatensätzen aus dem aufgenommenen Bestand: 30 je Evidenzklasse, innerhalb der Schicht zufällig gezogen, mit dokumentiertem Startwert. Personenbeziehbare Datensätze sind bauartbedingt ausgeschlossen, da nur bevölkerungsbezogene Datensätze in die Referenzmenge eingehen.

5.3 Menschliche Bewertung

Drei methodisch geschulte Bewerter, unabhängig arbeitend, verblindet gegenüber dem Maschinenwert. Je Datensatz zwei Merkmale: (a) fünfstufiges Sicherheitsurteil, (b) Paarvergleich innerhalb zufällig gebildeter Paare derselben Schicht. Die Übereinstimmung wird als Krippendorffs α mit Bootstrap-Konfidenzintervall berichtet. Bei $\alpha < 0,60$ wird das Erhebungsinstrument überarbeitet und die Runde wiederholt, bevor überhaupt mit der Maschine verglichen wird — eine Maschine lässt sich nicht gegen eine Referenz kalibrieren, die mit sich selbst uneins ist.

5.4 Primärer Endpunkt

Spearman-Korrelation zwischen EWR-Endwert und mittlerem menschlichem Sicherheitsurteil über alle 300 Datensätze.

- **Vorab festgelegter Erfolg:** $\rho \geq 0,70$ mit unterer Konfidenzgrenze $\geq 0,55$.
- **Vorab festgelegter Misserfolg:** $\rho < 0,50$ oder systematische Umkehrung in mehr als 15 % der klassenübergreifenden Paare, in denen Bewerter und Maschine in der Richtung auseinandergehen. Im Misserfallsfall wird die Klassengewichtstabelle aus den Bewertungsdaten neu aufgebaut und nicht nachgebessert.

5.5 Sekundäre Endpunkte

- Ablation: Werte allein aus dem Klassenrang gegen den vollen Modifikatorsatz. Verbessert das volle Modell ρ nicht um mindestens 0,05, trägt der Modifikatorapparat seine Komplexitätskosten nicht und wird reduziert.
- Alternative Verknüpfung: log-additive Abschlüsse mit gedeckelter Gesamtabwertung, geprüft an derselben Referenzmenge. Das bessere Modell gewinnt; geändert wird die Arbeit, nicht die Daten.
- Rückzugsbehandlung: eine eingestreute Menge von 15 bekannt zurückgezogenen Arbeiten muss in 15 von 15 Fällen mit 0,00 und Unsicherheit kritisch bewertet werden. Das ist ein Bestehen-oder-nicht-Kriterium, keine Korrelation.

5.6 Was die Kalibrierung nicht zeigen kann

Dass die Gewichte in einem absoluten Sinn richtig sind. Menschliche Bewerter sind eine Konvention, keine Grundwahrheit. Eine gelungene Kalibrierung erlaubt genau eine Aussage: Das System ordnet Datensätze annähernd so, wie geschulte Bewerter es täten. Alles Weitergehende verlangte Ergebnisdaten, die dieser Entwurf nicht erhebt und nicht erheben soll.

6. Datenqualität und Prüfung

6.1 Qualitätstor

Jeder aufgenommene Datensatz durchläuft vier Stufen: (1) Schemaprüfung (Pydantic v2), (2) Referenzprüfung gegen Fremdschlüssel und kontrollierte Vokabulare, (3) Dublettenerkennung (InChIKey / PMID / NCT-ID / DOI), (4) Plausibilität ($IC_{50} > 0$, $p \in [0,1]$, Molmasse > 0 , SMILES über RDKit lesbar, LOINC-Code auflösbar). Fehlschläge gehen mit protokollierter Regel in Quarantäne und bleiben abfragbar. Das Löschen eines Quarantänedatensatzes ist eine ausdrückliche Kuratorenhandlung mit Protokolleintrag.

6.2 Automatisierte Tests

51 Tests: Evidenzmodul 20, Qualitätstor 7, Laborparser 9, Biomarkerstatus 4, Hypothesenintegration 3, DSGVO-Vorgänge 7 (Art.-15-Export, Art.-17-Löschung, Einwilligungszustandsmaschine), API 15. Die Zählungen überschneiden sich dort, wo ein Test zwei Bereiche abdeckt; die Testkennungen stehen im Validierungsprotokoll.

Was die Tests nicht abdecken — hier ausgeschrieben, weil die Formulierung aus v0.1 („validated with 51 automated tests“) zu weit ging: Sie zeigen, dass die Umsetzung der Spezifikation entspricht. Sie zeigen nichts darüber, ob die Spezifikation richtig ist. Die Bewertungsgüte im Sinne von §5 ist ungeprüft. Connectorverhalten gegen Live-Schnittstellen unter Ratenbegrenzung, Schemadrift bei Vorlieferanten und Langzeitkonsistenz des Graphen sind nicht abgedeckt und als L9–L11 geführt.

6.3 Reproduzierbarkeit

Deterministische Bewertungsfunktionen ohne Zufall; versionierte Bewertungskonfiguration; Konfigurationsversion auf jedem Wert; fortschreibendes Prüfprotokoll für Bewertungsänderungen; Containerabbilder über Digest festgenagelt. Eine Neuberechnung des Gesamtbestands unter einer gegebenen Konfigurationsversion muss jeden gespeicherten Wert bitgleich reproduzieren; das ist selbst ein Test (T-REP-01) und läuft vor jeder Freigabe.

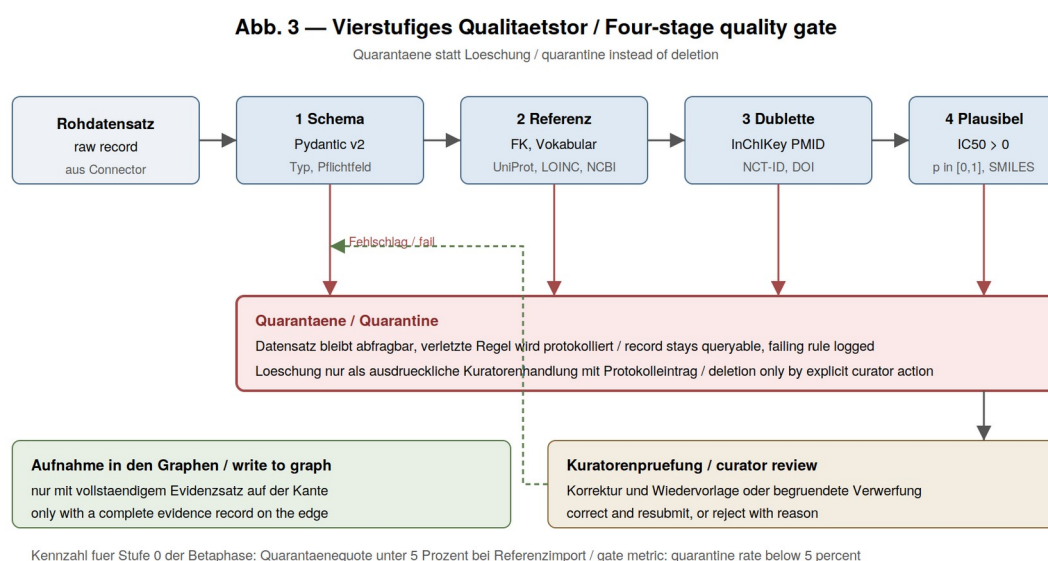


Abbildung 3 — Qualitätstor. Abgewiesene Datensätze bleiben abfragbar.

7. Regulatorische Einordnung

Dieser Abschnitt ersetzt den einzelnen Haftungsausschluss aus v0.1. Er ist eine Einordnung, keine Rechtsberatung; der Autor ist kein Jurist, und die mit [R] markierten Punkte verlangen anwaltliche Prüfung vor jeder Freigabe über einen geschlossenen Testkreis hinaus.

7.1 Medizinprodukte-Qualifikation (MDR/IVDR)

Die Qualifikation folgt der Zweckbestimmung, nicht der Technik, und nach MDCG 2019-11 Rev. 1 (Juni 2025) wird ein mehrmoduliges Produkt modulweise bewertet: Ein medizinisches Modul innerhalb einer sonst nicht-medizinischen Plattform zieht nur dieses Modul in das Produktregime, sofern Grenzen, Schnittstellen und Abhängigkeiten dokumentiert sind [16]. Die

Überarbeitung führt zugleich den Begriff der medizinischen KI ein und verweist auf den Europäischen Gesundheitsdatenraum — deshalb legt §3.5 die Modulgrenzen ausdrücklich fest.

Angewandt: M1–M3, M6 und M7 verfolgen einen Forschungszweck und liefern keine Information zur Diagnose oder Behandlung einer Krankheit. M5 (Biomarker-Kontext) ist die Grenze. In der spezifizierten Form — Anzeige eines Werts gegen den Referenzbereich des Labors, ohne Deutung, ohne Risikowert, ohne Anstoß zu weiterer Diagnostik — qualifiziert M5 nicht. Folgende Änderungen kippen die Einordnung:

Auslöser	Wirkung
Deutung eines Werts als Hinweis auf einen Zustand	Qualifikation als MDSW; Klassifizierung nach Anhang VIII Regel 11 [R]
Erzeugung eines Krankheitsrisikowerts	Qualifikation; voraussichtlich Klasse IIa oder höher [R]
Empfehlung weiterer Diagnostik	Qualifikation (das MDCG-Beispiel zum Risikowert ist unmittelbar vergleichbar) [R]
Verdichtung mehrerer Werte zu einem als Gesundheitsinformation dargestellten Index	Qualifikation [R]
Werbeaussage über Erkennung, Verhütung oder Überwachung von Krankheit	Qualifikation unabhängig vom technischen Entwurf

Die letzte Zeile verdient die Betonung: Die Zweckbestimmung entsteht aus dem, was behauptet wird, einschließlich in Werbematerial. Ein Haftungsausschluss heilt keine Aussage, die anderswo auf derselben Seite steht. Flyer und Netzseite dieses Bündels sind gegen diese Vorgabe geschrieben.

7.2 Regulierung künstlicher Intelligenz

Zwei Fragen, in dieser Reihenfolge.

Ist es ein KI-System? Das Bewertungsmodul ist eine deterministische Regelauswertung mit festen Koeffizienten. Es leitet aus Eingaben nicht ab, wie Ausgaben zu erzeugen sind, und passt sich nicht an. Auf dieser Grundlage fällt es nicht unter die Begriffsbestimmung in Art. 3 Abs. 1 der Verordnung (EU) 2024/1689. Das gilt allein für die laufende Version; die auf v2 verschobene NLP-Extraktion (L5) würde die Antwort für dieses Modul ändern.

Und falls doch, was gilt? Art. 2 Abs. 6 nimmt KI-Systeme und -Modelle aus, die allein für wissenschaftliche Forschung und Entwicklung entwickelt und in Betrieb genommen werden. Diese Ausnahme deckt die laufende Nutzung und endet dort, wo das System in Verkehr gebracht oder für nicht-forschende Zwecke in Betrieb genommen wird — genau der Übergang, auf den das Betaprogramm in §8 zuläuft.

Zeitplan nach der Änderung von 2026: allgemeine Anwendbarkeit ab 2. August 2026 einschließlich der Transparenzpflichten nach Art. 50 für KI-erzeugte Inhalte und für Interaktion mit Menschen. Die Pflichten für eigenständige Hochrisikosysteme nach Anhang III wurden durch den Digital Omnibus vom 2. August 2026 auf den 2. Dezember 2027 verschoben, für in regulierte Produkte eingebettete KI nach Anhang I auf den 2. August 2028. Weitere

Verbote gelten ab 2. Dezember 2026. [R] für die konkrete Wirkung auf ein künftiges v2-NLP-Modul, denn eine verschobene Pflicht ist keine aufgehobene.

7.3 Datenschutz

Biomarkerwerte sind Gesundheitsdaten nach Art. 9 Abs. 1 DSGVO. Rechtsgrundlage ist die ausdrückliche Einwilligung nach Art. 9 Abs. 2 lit. a, je Verarbeitungszweck getrennt, widerruflich, mit dem Widerrufsweg in der Oberfläche statt in einem Erklärungstext. Eine Datenschutz-Folgenabschätzung nach Art. 35 liegt im Entwurf vor und ist prüfungsbedürftig [R]. Betroffenenrechte sind als API-Vorgänge mit Tests umgesetzt (Art.-15-Export, Art.-17-Löschung, Art.-20-Übertragbarkeit). Die Speicherbegrenzung ist je Datenkategorie konfiguriert. Die Pseudonymisierung trennt Identitäts- und Wertetabelle; das mindert den Personenbezug, hebt ihn nicht auf, und die Arbeit behauptet keine Anonymisierung.

Für einen unabhängigen Forscher ohne institutionelle Ethikkommission ist der Ethikweg vor der ersten externen Testrunde zu wählen: entweder Votum eines unabhängigen Ethikgremiums oder eine dokumentierte, schriftliche Begründung, warum keines erforderlich ist, abgelegt vor der ersten Teilnahme.

7.4 Europäischer Gesundheitsdatenraum

Die Verordnung (EU) 2025/327 ist am 26. März 2025 in Kraft getreten und gilt gestaffelt: allgemeine Anwendbarkeit ab 26. März 2027, das Kapitel zur Sekundärnutzung — Gesundheitsdatenzugangsstellen, Datengenehmigungen, HealthData@EU — ab 26. März 2029, weitere Datenkategorien ab 2031. Zwei Folgen für diesen Entwurf: Ein künftiger Weg zu bevölkerungsbezogenen Gesundheitsdaten führt über eine Zugangsstelle und eine Datengenehmigung, nicht über bilaterale Absprachen; und die Interoperabilitätsanforderungen an elektronische Patientenaktensysteme bilden die Bezugsarchitektur für den Laborimport, weshalb die LOINC-Codierung in M4 verpflichtend und nicht optional ist.

7.5 Zusammenfassung der Einordnung

Regelwerk	Stand	Auslöser für Änderung
MDR/IVDR	kein Medizinprodukt; M5 ist die Grenze	Deutung, Risikowerte, Diagnostikanstoß, Krankheitsaussagen in der Werbung
KI-Verordnung	außerhalb Art. 3 Abs. 1; Forschungsausnahme Art. 2 Abs. 6	v2-NLP-Modul; Inverkehrbringen über Forschung hinaus
DSGVO	Einwilligung Art. 9 Abs. 2 lit. a, DSFA im Entwurf [R]	jeder Zweck außerhalb des Einwilligungstexts
EHDS	vor 2027/2029 nicht einschlägig	jede Nutzung von Gesundheitsdaten Dritter
Österreichisches Recht	DSG, Offenlegung nach § 5 ECG für die öffentliche Seite	gewerblicher Betrieb, Werbung

Abb. 4 — Regulatorischer Entscheidungspfad mit Kippunkten / Regulatory decision path with triggers

Einordnung, keine Rechtsberatung — mit [R] markierte Punkte sind prüfungsbedürftig / positioning, not legal advice

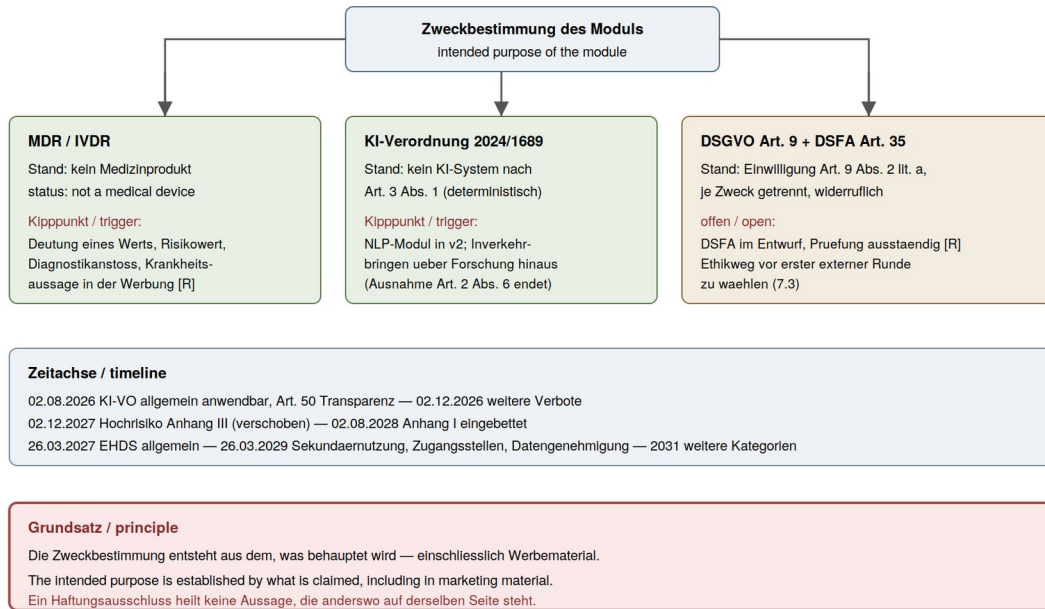


Abbildung 4 — Regulatorische Einordnung mit Kippunkten und Zeitachse. Mit [R] markierte Punkte sind prüfungsbedürftig.

8. Prüfung und Betaprogramm

8.1 Zweck

Geprüft werden Bedienbarkeit, Verständlichkeit der Unsicherheitskennzeichnung und Robustheit der Connectoren — nicht die Gesundheitswirkung von irgendetwas. Es wird kein Gesundheitsendpunkt erhoben, keine Maßnahme verabreicht, kein Rat erteilt.

8.2 Stufen

Stufe 0 — intern. Nur synthetische Daten. Ausstiegskriterium: alle 51 Tests grün, T-REP-01 reproduzierbar, Quarantänequote unter 5 % bei einem Referenzimport.

Stufe 1 — geschlossene Alpha, 5–8 Teilnehmende, keine personenbezogenen Gesundheitsdaten. Nur bevölkerungsbezogene Funktionen. Endpunkt: Können Teilnehmende allein aus der Oberfläche richtig angeben, was eine Unsicherheitskennzeichnung bedeutet? Erhebung über zehn Verständnisfragen; vorab festgelegte Schwelle 80 % richtig.

Stufe 2 — geschlossene Beta, 15–25 Teilnehmende, eigene Laborbefunde, ausdrückliche Einwilligung. Beginn nicht vor: geprüfter DSFA [R], geklärtem Ethikweg (§7.3), von dritter Seite durchgetestetem Löschweg und abgeschlossener Kalibrierung nach §5 — andernfalls trägt jede Wertanzeige die Kennzeichnung „unkalibrierte Ordnung“.

Stufe 3 — offene Beta. Erst nach Erfüllung der Abbruchkriterien aus Stufe 2 und nach Prüfung der regulatorischen Einordnung [R].

8.3 Abbruchregeln

Jeder der folgenden Punkte hält das Programm an und löst einen schriftlichen Vorfallsbericht aus: ein unbehobener Export- oder Löschfehler; die Meldung einer teilnehmenden Person, sie habe aufgrund einer Anzeige eine Gesundheitsentscheidung getroffen; ein Connector, der Daten der falschen Quelle zuschreibt; ein Wert, dessen Herleitung sich aus dem gespeicherten `applied_modifiers[]` nicht rekonstruieren lässt.

8.4 Datenminimierung

Teilnehmende laden nur hoch, was sie hochladen wollen. Keine Kontenzusammenführung, keine Drittanalyse, keine externen Schriften oder Skripte auf der öffentlichen Seite. Rückmeldungen werden in einer Form erhoben, die keine Wiederholung von Gesundheitsdaten verlangt.

9. Einschränkungen

L1 — Keine empirische Kalibrierung. Die Gewichte sind begründet, nicht gemessen. §5 ist ein Plan, kein Ergebnis. Bis zur Durchführung ist kein Wert als Wahrscheinlichkeit oder Sicherheitsschätzung zu lesen.

L2 — Keine Kausalaussagen. Ausschließlich Assoziationen, bei jedem Wert.

L3 — Assay-Unvergleichbarkeit. IC50-Werte aus verschiedenen Testsystemen sind nicht unmittelbar vergleichbar. Assay-Metadaten werden annotiert; eine assayübergreifende Normalisierung findet nicht statt und ist ohne validiertes Verfahren nicht geplant.

L4 — Taxonomische Instabilität. Ältere Arbeiten verwenden überholte Pflanzennamen. Die Vereinheitlichung über die NCBI-Taxonomie löst nicht alle historischen Synonyme auf; ungelöste Namen werden gekennzeichnet, nicht geraten.

L5 — NLP verschoben. Automatische Volltextextraktion ist wegen des Falsch-Positiv-Risikos und wegen der Wirkung auf die KI-Einordnung (§7.2) auf v2 verschoben. v1 nutzt ausschließlich strukturierte Importe.

L6 — Rechtsprüfung ausständig. Alle mit [R] markierten Punkte in §7 verlangen fachkundige Prüfung vor jeder Freigabe über einen geschlossenen Kreis hinaus.

L7 — Unabhängigkeit der Modifikatoren unterstellt. Das multiplikative Modell behandelt Modifikatoren als unabhängig; sie sind es nicht (§4.7). Das Alternativmodell ist in §5.5 vorab festgelegt.

L8 — Keine Pfadunsicherheit. Mehrschrittige Traversierungen melden die schwächste Kante. Konservativ und ungenau.

L9 — Schemadrift bei Vorlieferanten. Connectoren sind gegen Fixtures getestet, nicht gegen Live-Änderungen. Eine stille Feldänderung stromaufwärts erschien derzeit als Quarantäneausschlag, nicht als benannter Fehler. Ein belegter Fall liegt vor: die klassische Schnittstelle von ClinicalTrials.gov wurde am 25. Juni 2024 abgeschaltet, verbindlich ist API v2 mit anderer Basisadresse und anderer Blätterung; Code gegen die alte Fassung bricht. Der Studien-Connector ist entsprechend nachzuziehen.

L10 — Quellenprüfung unvollständig. Von 55 registrierten Quellen sind 18 geprüft (32,7 %, Stand 2026-09-04, zwei Prüfrunden). Die übrigen 35 stammen aus v0.1 und dürfen nicht als geprüft zitiert werden. Die zweite Runde ergab bei sechs Prüfungen vier Änderungen — darunter eine abgeschaltete Schnittstelle und eine Lizenzbedingung, die auf einen Forscher ohne institutionelle Anbindung nicht zutrifft (KEGG). Im ungeprüften Rest ist mit weiteren Fehlannahmen zu rechnen.

L11 — Einzelbearbeiterrisiko. Eine Person baut, prüft und gibt frei. Es gibt keine unabhängige Codeprüfung. Das ist genannt, weil es die Belastbarkeit jeder „geprüft“-Aussage in §6 begrenzt.

L12 — Referenzmenge nicht gezogen. Die 300 Datensätze aus §5.2 existieren noch nicht. Das Protokoll ist veröffentlichungsfähig, das Ergebnis nicht, und diese Version behauptet keines.

10. Schluss und Weg zu v1.0

Das System führt Naturstoffevidenz und individuelle Biomarkerwerte unter einer Gewichtung- und Unsicherheitssemantik zusammen, hält Provenienz auf jeder Kante und verweigert zwei bestimmte Übertragungen (Extrakt—Reinsubstanz, Tier—Mensch) bereits beim Schreiben. Version 0.3 nimmt die GRADE-Anpassungsbehauptung zurück, benennt, was die Zahlen nicht bedeuten, und ergänzt das Kalibrierungsprotokoll, das den Bewertungsentwurf widerlegen kann.

Weg zu Version 1.0, in dieser Reihenfolge und mit Toren:

1. Kalibrierung nach §5 durchführen; Ergebnis veröffentlichen, wie immer es ausfällt.
2. Rechtsprüfung aller [R]-Punkte; DSFA abschließen.
3. Verständnisstudie Stufe 1; Kennzeichnungen überarbeiten, falls unter Schwelle.
4. Unabhängige Prüfung des Bewertungs-codes durch eine zweite Person.
5. Quellenregister auf 100 % geprüft bringen oder den ungeprüften Rest entfernen.
6. Zeitschrifteneinreichung auf Schwelle B: Primärziel *Journal of Cheminformatics* (Software-Artikel; verlangt vollständige Reproduzierbarkeit und Quellcode unter OSI-Lizenz; Publikationsgebühr fällt an), Alternativen *Database* (Oxford) oder *JAMIA Open*; die Softwarekomponente zusätzlich an *JOSS*, das keine Gebühr erhebt und die Software statt eines Fließtexts begutachtet.

Version 1.0 ist erreicht, wenn 1–5 abgeschlossen sind, nicht wenn der Code fertig ist.

Quellen

Die Quellenliste ist mit der englischen Fassung identisch; Nummerierung [1]–[28] ist über beide Sprachfassungen hinweg gleich. Siehe FORSCH-PROV0427_biologic-fbs-paper_paper_v0.3_2026-09-04_EN.md, Abschnitt „References“, sowie das Quellenregister FORSCH-PROV0427_biologic-fbs-paper_quellenregister_v0.3_2026-09-04_DE-EN.md.

Anhang A — Abbildungen

- **Abbildung 1** — Systemarchitektur mit Modulgrenzen (M1–M7) und der regulatorischen Bruchlinie an M5.
- **Abbildung 2** — Evidenzgewichtung: Klassengewicht, Modifikatorkaskade, Deckelung, abgeleitete Unsicherheit.
- **Abbildung 3** — Vierstufiges Qualitätstor mit Quarantäneschleife und Kuratorenhandlung.
- **Abbildung 4** — Regulatorischer Entscheidungspfad mit Kipppunkten (MDR/IVDR, KI-VO, DSGVO, EHDS).

Anhang B — Prüfstand der Quellenbasis

Von 28 Quellen wurden 14 am 2026-09-04 gegen den Verlagsnachweis nachgeprüft (DOI, Band, Seiten, Jahr): [1], [4], [5], [9], [10], [11], [12], [13], [16], [21], [22], [23], [25], [26]. Die übrigen sind fortgeschriebene Standardwerke und Rechtstexte; Rechtstexte sind über ihre amtliche Kennung zitiert und wurden dort auf den Änderungsstand geprüft, wo der Änderungsstand für das Argument zählt ([21], [22]). Quelle [15] ist eine Datenressource, keine Veröffentlichung, und wird über das Zugriffsdatum zitiert.

Fußzeile: Thomas Schüller · Syntagma Forschung, Wien · forschung@tschueller.com · ORCID 0009-0003-9799-6747 · CC BY 4.0 · v0.3 · 2026-09-04 · FORSCH-PROV ≥0427 (provisorisch) · Slug biologic-fbs-paper