RESPONSE TO PUBLIC CONSULTATION

Anonymising clinical narrative: a comment from measured practice

On Guidelines 02/2026 on Anonymisation, version 1.0, adopted 7 July 2026.

RESPONDENT	Filippos-Paschalis Rorris
AFFILIATION	Department of Pediatric and Adult Congenital Cardiac Surgery, Onassis Hospital, Athens, Greece
CONTRIBUTOR TYPE	Individual
BASIS OF COMMENT	Development and published benchmarking of a clinical-text de-identification engine for Greek

The Guidelines are general and sector-neutral but comments are drawn from one domain — clinical narrative — and are offered as a concrete test of how the general framework applies to it, not as a request for special treatment of health data. Where we describe a difficulty, we have tried to indicate whether it is particular to medical records or a property of unstructured text as such.

The comments come from building and measuring a de-identification system for clinical text in a language that had no annotated corpus, and are offered in the spirit of paragraph 84, which encourages the development and dissemination of scientifically grounded expertise. Five comments follow, each tied to the paragraphs it concerns.

Comment 1

The technical framework is defined for structured data, and clinical narrative falls outside it

Concerning paragraphs 2, 44 and 85 of the Guidelines

The three criteria and the factors offered for applying them are expressed in the vocabulary of tabular, record-level data. Paragraph 85 lists among the key factors *"the dimensionality of the data … the number of columns in a table for each individual record"*, together with resolution and whether data is aggregated or at record level.

Clinical documents vary greatly in format, and the difficulty is not that they are wholly unlike tables. Laboratory and imaging reports in particular carry genuinely tabular sections — analyte, value, reference range — and for those sections the guidance's factors apply as they would to any table. But the substance of a discharge summary, an operative note or a referral letter is prose, and most reports pair whatever table they contain with narrative interpretation. Identifiers appear in both parts.

It is the narrative part that the factors cannot reach. Prose has no countable dimensionality, no record-level granularity in the sense used here, and no aggregation state, so a controller assessing it has nothing in paragraph 85 to apply. The difficulty is not peculiar to medicine: the same is true of a legal opinion or a social-care case note. It is a property of unstructured text, and it is where the identifiers that matter most — names, addresses, the incidental detail of a clinical history — actually sit.

This is not a marginal case. The introduction at paragraph 2 gives the establishment of EU-wide data spaces as one of the reasons the 2014 Opinion required updating; the guidelines then do not address health data, and the words health, clinical, unstructured and narrative do not appear in the text. Yet under Regulation (EU) 2025/327 the secondary-use chapter applies from 26 March 2029, and a substantial share of what data holders will be asked to make available is narrative: discharge summaries, operative notes, imaging and pathology reports, referral letters.

> **Comment to the Board.** We invite the Board to state, even briefly, how the three criteria are to be applied to unstructured text, or to acknowledge in paragraph 44 that this document does not yet cover it. Either would help controllers more than silence, which currently leaves each of them to improvise a mapping from columns to prose.

Comment 2

What documented effectiveness can look like, from a worked example

Concerning paragraphs 41 and 84 of the Guidelines

Paragraph 41 requires documentation of the anonymisation process *"including for the testing of the supposedly anonymous datasets"*, sufficient to demonstrate both compliance and the effectiveness of the anonymisation. We support this requirement, and offer our own practice as one concrete answer to what can satisfy it.

We constructed a held-out benchmark of 500 synthetic Greek clinical discharge notes. Every identifier was pre-annotated at generation and then reviewed and corrected, across the whole corpus, by two clinicians working independently and blinded to each other, with disagreements resolved by arbitration. Each note also carries clinical items — drug names, diagnoses, laboratory values, eponyms — that must survive de-identification intact. Every item thus carries one of two expectations: must be removed where it is an identifier, or must survive where it is clinical content. No component of the evaluated system was tuned against this corpus.

MEASURE	RESULT	DETAIL
Identifier removal (recall)	**96.59%**	7,666 / 7,937 · 95% CI 96.16–96.96
Direct identifiers	**98.49%**	95% CI 98.11–98.79
Clinical content destroyed	**2.82%**	190 / 6,747 must-survive probes
Inter-annotator agreement	**F1 0.998**	pre-arbitration

Measured on the frozen engine, 500 notes.

Two features of this are worth drawing out, because they are what make the documentation meaningful rather than declarative.

EFFECTIVENESS HAS TWO DIRECTIONS, NOT ONE

A de-identification rate reported alone is not interpretable, because its maximum is trivially available: a system that deletes the document scores one hundred per cent. We therefore treat the destruction of clinical content as a co-primary outcome, measured against planted probes. The field does not generally report this. The authors of one of the few externally certified deployed systems noted that no standard existed for measuring the information lost to false positives; that remains broadly true, and it means a controller reading a vendor's recall figure learns only half of what paragraph 41 asks them to document.

THE PROCESS DOCUMENTED MUST BE THE PROCESS RUNNING

Paragraph 41 presumes that the pipeline whose effectiveness was documented is the pipeline in production. We make that checkable rather than assumed: model revisions and runtime settings are pinned, every processed document is stamped with a fingerprint of the exact configuration that handled it, and a frozen probe document with a known expected output is verified at start-up, with processing halted if the engine's behaviour has changed. Running the full benchmark twice in independent launches produced identical verdicts on all 6,747 clinical items and outcome measures equal to within rounding; the 65 of 14,661 scored decisions that differed (0.44%) all arose from the deliberately randomised replacement text, and none from a decision of the pipeline.

> **Comment to the Board.** We invite the Board to indicate that quantitative, published benchmarking satisfies paragraph 41, and to make explicit that reproducibility and auditability of the process — the ability to show that the configuration validated is the configuration that ran, and to re-derive a past decision — form part of what documented effectiveness means. As drafted, a controller could satisfy paragraph 41 with a description of a process no one can verify was the one applied.

Comment 3

What we cannot currently measure, offered as a gap in the field rather than a solved problem

Concerning section 3.4 and paragraph 84 of the Guidelines

We would not want the figures above to imply more than they support. Identifier recall and content preservation speak to whether identifiers were removed. They do not measure singling out or inference on the residual narrative, and we are not aware of an established method that does for free text.

This matters for the framework directly. A clinical narrative can fail No Record Isolation or No Inference with every name correctly removed: a rare diagnosis, an institution and an approximate date may together isolate one person, and nothing in an identifier-recall figure would reveal it. Structured data has k-anonymity and its successors for this; narrative text has, so far as we know, nothing equivalent that a controller could apply and document.

> **Comment to the Board.** We raise this as an open problem rather than a criticism. If the Board expects controllers to evidence criteria 1 and 3 on unstructured text, it would help to say what form that evidence could take; if no such method is yet established, saying so would itself be useful, and would direct research effort where paragraph 84 intends.

Comment 4

The guidance treats AI models only as a threat, and is silent on the instrument that performs the anonymisation

Concerning paragraphs 77, 82, 83 and 92–93 of the Guidelines

Generative models appear throughout as attackers: extraction attacks against supposedly anonymous models (¶77), inference drawn by prompting a model or synthetic data (¶82), special-purpose agents deploying combinations of probabilistic techniques (¶83), and agentic AI lowering the time and cost of re-identification (¶92–93). We think this analysis is right. It is also incomplete, because the same technology is increasingly proposed as the tool that performs anonymisation, and the guidelines say nothing about that side.

Our position, offered as practitioners rather than as a request for prohibition: a generative model is the wrong instrument for the de-identification step itself, for reasons that follow from the Board's own framework.

- **Auditability.** Paragraph 41 presumes the validated process is the running process. A pinned, fingerprinted, self-testing pipeline can demonstrate that; a sampling generative model cannot, and a validation performed last year need not describe this year's behaviour.
- **Failure mode.** A span-based system either removes a span or leaves it, and its errors are countable, span by span, against a gold standard. A generative system that rewrites the document can fail by fluent fabrication — an altered dose or date that reads perfectly well. Content corruption then becomes unbounded and, in practice, unmeasured.
- **Data minimisation.** The de-identifier sees the record before any protection exists. Routing identifiable records to a hosted model concentrates unprotected data at the least transparent point in the chain.
- **The model itself.** On the Board's own reasoning at ¶77 and ¶82, and in Opinion 28/2024, a model trained on clinical records may itself hold personal data. A model fine-tuned on real notes to learn de-identification is a new repository of patient data created before a single document has been protected.
- **It is not necessary.** Deterministic, non-generative pipelines reach the accuracy the task requires: our own Greek figures above, and in English a rule-driven, offline, externally certified system reporting token-level recall of 99.5%.

We would note for completeness that our engine is not free of machine learning — it uses discriminative recognisers for named entities. The distinction we draw is between deterministic, non-generative components, whose decisions can be frozen, fingerprinted and replayed, and generative models, whose output cannot.

> **Comment to the Board.** We do not ask the Board to endorse or exclude any technology. We ask only that the reproducibility and auditability requirement in comment 2 be made explicit, since it is technology-neutral on its face while being decisive in practice: it can be met by a system whose decisions are replayable, and cannot be met by one whose decisions are not.

Comment 5

Two closing points: terminology, and the generality of the method

Concerning paragraphs 40 and 84 of the Guidelines

ON TERMINOLOGY, AS IT APPLIES TO OUR OWN SYSTEM

Paragraph 40 cautions against describing data as anonymous or de-identified where individuals remain identifiable. We accept the point as it applies to our own system: where a hospital retains the mapping between real and replacement values, the outgoing text is anonymous relative to the recipient and pseudonymous for the hospital; the unqualified term, where it is used for brevity, should be read in that sense. We mention this only because the relative assessment the guidelines describe is easy to endorse and harder to write down honestly, and the clarity of paragraph 40 helped.

ON GENERALITY

Although this system was built for clinical text, nothing in the approach is health-specific. Layered recognisers, curated vocabularies, surrogate replacement, an independent scan of the outgoing text, and a determinism perimeter that makes decisions reproducible could be tuned to

any document class where anonymisation is required — legal, administrative, social-services records. Clinical narrative is the proving ground, being among the densest in identifiers and the highest in consequence, not the boundary of the method. We offer the practice in that spirit: as a worked example a general framework could accommodate, not a request for a sectoral exception.

Supporting material

The benchmark corpus, the arbitrated gold standard, the complete scoring implementation and the evaluated system's per-decision outputs are published so that every figure above can be independently recomputed: https://doi.org/10.5281/zenodo.22824894. The accompanying study (Rorris F-P, Stanitsa N, Rorri T, Rorri F, Antonopoulos CN. Morpheus: a fully local, defense-in-depth de-identification engine for Greek clinical text — a benchmark evaluation) was submitted to a peer-reviewed journal on 1 October 2026 and is under review. We are glad to provide further detail if it would assist the Board.