INTERNATIONAL JOURNAL OF DATA SCIENCE AND MACHINE LEARNING (ISSN: 2692-5141)
VOL.06 ISSUE 02 (2026)

Research Article

Risk-Weighted Evaluation of De-Identification Tools on Synthetic Clinical Trial Narratives

Vanshita Moolchandani
Independent Researcher, Boston, MA, USA

Received: 06 November 2026
Revised: 15 November 2026
Accepted: 16 November 2026
Published: 27 November 2026

Doi: 10.55640/ijdsml-06-02-20
Page No: 313-334

Copyright: © 2026 Authors retain the copyright of their manuscripts, and all Open Access articles are disseminated under the terms of the Creative Commons Attribution License 4.0 (CC-BY), which licenses unrestricted use, distribution, and reproduction in any medium, provided that the original work is appropriately cited.

Abstract
De-identification software used to process clinical trial narratives prior to sharing them with regulators and the public is typically evaluated using F1, which treats all missed items equally. However, a missed rare diagnosis poses a greater threat to an individual's privacy than a missed common symptom. Therefore, we investigated whether four de-identification software packages would rank differently when evaluated using flat F1 versus a risk-weighted F1. This is an evaluation of the scoring technique on artificial text, and not an assessment of whether the weights correlate with actual re-identification risk. We generated 150 simulated narratives for three trials (oncology, rare disease, and psychiatry) from a fictional population of 10,000 individuals. Each item in the answer key has a risk tier and a weight (S3) based on how many people share that single value. S3 is a rarity proxy built from the frequency of individual attributes, and it does not estimate the re-identification risk of a whole narrative. Flat F1 here combines character-level precision with item-level recall, so it differs from the usual span-level F1. We evaluated Presidio, a disease and chemical named-entity recognition (NER) baseline (scispaCy NER), the RoBERTa de-identification model obi/deid_roberta_i2b2, and the language model qwen3:8b on 135 test narratives using strict matching and tested the rank changes with a bootstrap and the Holm correction. The language model performed best under both flat F1 (0.645) and risk-weighted F1 (0.813). The only rank swap we observed was between Presidio and the RoBERTa model. It was not a statistically confirmed reversal, because the bootstrap intervals for both differences include zero (Holm-adjusted p = 0.709 for flat F1 and 0.208 for risk-weighted F1). H1 was not supported. Although no reversal was confirmed, the weights did provide additional insight. The weights increased the performance of the language model relative to the other three tools, and they demonstrated that none of the tools identified more than 8.8% of occupations, unusual procedures, relatives' conditions, or gene variants. This study provides a methodology for weighting errors based on risk and a synthetic data set with an answer key for evaluating de-identification tools.

Keywords: clinical trial narratives; de-identification; re-identification risk; risk-weighted evaluation; synthetic data; large language models; clinical study reports

1. INTRODUCTION

Sponsors of clinical trials are now sharing more of their results. The European Medicines Agency (EMA) publishes clinical study reports under its Policy 0070, and Health Canada releases clinical information to the public under its own guidance [1], [2]. These reports contain patient narratives, which describe single participants who experienced a serious adverse event or died. A narrative can contain a rare diagnosis, a date, a family history, or an occupation, and any of these can identify a single person. Both regulators allow for a re-identification risk of 0.09, which Health Canada equates to a group of 11 patients [2]. In simple terms, a participant's details should be shared by at least 11 people, because 1 divided by 11 is about 0.09.

Additionally, the EMA guidance requires sponsors to retain adverse event terms in the text [1]. Manual de-identification of narratives is time-consuming, so sponsors consider automatic tools. These tools are typically evaluated using precision, recall, and F1, as in the i2b2 shared tasks [3], [4]. F1 treats all misses equally. A missed name and a missed mention of hypertension are treated equally, even though only the first identifies a person. A recent scoping review of 110 clinical text de-identification studies highlights this issue: two systems with similar F1 may differ significantly in the high-risk identifiers they miss. The review recommends additional evaluations, such as audits of residual identifiers and adversarial re-identification tests, but does not suggest a risk-weighted score [5]. Over-removal creates a second problem. A survey of LLM-based de-identification indicates that standard classification metrics fail to capture errors that alter clinical content [6]. Free text also contains quasi-identifiers that are not covered by any list of identifier types. Diagnosis codes alone can identify research participants [7], and one study evaluated the re-identification risk of a clinical study report de-identified according to EMA Policy 0070 and Health Canada rules [8]. Prior research has linked evaluation to risk in two ways. Scaiano et al. developed a framework for evaluating the risk of re-identification of text de-identification tools [9]. Graff et al. evaluated NLP tools for de-identifying clinical trial narratives and argued that standard metrics are insufficient because some entities are more important to remove than others [10]. The same research is presented in Chapter 6 of Graff's thesis [11], which establishes a framework for weighting items differently based on intruder type (Table 6.5 of the thesis). The table describes intruder types and what each may know about a participant.

However, it does not include numeric weights or a formula, so a tool cannot be evaluated using it. This leaves two gaps. To our knowledge, no evaluation method applies graded weights to every item in a narrative, including medical details.

Furthermore, no shareable data set exists that allows researchers to evaluate such an evaluation method because real narratives cannot be shared. This study addresses one research question (RQ1): on synthetic clinical trial narratives, do de-identification tools rank in a different order using F1 than using a risk-weighted score? We establish three objectives. The first is to generate synthetic narratives with a known answer key. The second is to assign each item a rarity-based weight (S3), taken from how many people in the synthetic population share that item's value, and to use the 0.09 threshold to set the risk tiers. The third is to compare four tools using flat and risk-weighted F1, using a bootstrap test that accounts for chance. We build upon Scaiano et al. [9] by applying graded, risk-based weights to every item, including medical details, and demonstrate on shareable synthetic trial narratives how much the choice of evaluation method affects the difference between tools and which pairs of tools swap positions. Our hypothesis (H1) was that at least one pair of tools would swap positions in a manner that a test could confirm. H1 was not supported. The main contribution is a scoring method: a risk tier and a rarity-based weight for every item, taken from the number of people who share that single value, and a weighted F1 built on those weights. It does not propose a new taxonomy of identifiers, a tool architecture, or a rule for choosing between tools. Two further products come with the method. One is a way to generate tagged synthetic narratives together with an answer key, which other groups can reuse and which makes the counts behind the weights exact. The other is a comparison of four tools showing that the choice of score changes how far apart they appear, and that all four miss more than 90% of occupations, unusual procedures, relatives' conditions and gene variants.

The S3 weight needs a plain description. For one item, S3 is 1 divided by the number of people in the synthetic population who share that item's value. It is a rarity proxy: it tells us how unusual a single value is, and it comes from the frequency of individual attributes. The 0.09 threshold uses the same arithmetic, 1 divided by the size of a group of people who look alike, so we used it to place the tier boundaries. The threshold, however, applies to the combination of everything an intruder knows about a person, and S3 looks at one attribute at a time. A narrative holds 11 to 19 items, and the people who match all of them together cannot be more numerous than the people who match the rarest one. S3 therefore does not estimate the re-identification risk of a whole narrative, and RW-F1 is not a probability of re-identification. It counts a missed rare item as a bigger miss than a missed common one. The Methodology gives the limits in detail.

Literature Review

Early research evaluated de-identification using recall and precision on annotated corpora, as in the i2b2 shared tasks [3], [4]. Meystre et al. reviewed the research on automatic de-identification of health records [12], and Ferrández et al. evaluated existing methods on Veterans Health Administration documents [13]. Two earlier frameworks link evaluation to re-identification risk, and ours is closest to them. Scaiano et al. developed a unified framework for evaluating the risk of re-identification of text de-identification tools [9]. Their framework works at the level of identifier types. Graff et al. argued that standard metrics are not enough because some entities matter more to remove than others, and their thesis outlines how weights could differ by intruder type [10], [11]. It gives no numeric weights or formula, so a tool cannot be scored with it. Our scheme takes the same starting point and differs in two ways. It gives a numeric weight to every individual item, including medical details such as rare diagnoses and unusual procedures, and that weight comes from a count of how many people share the item.

A second line of research examines what tools remove that they should not. Aghakasiri et al. surveyed LLM-based de-identification and argued that standard classification metrics fail to capture errors that alter clinical content. Clinical experts validated an existing metric for this loss, and the authors developed a new method for identifying the removal of clinically relevant information [6]. Sariyar reviewed 110 studies and recommended measures beyond F1, such as analysis of high-risk residual errors, utility assessment, and adversarial testing [5]. The Text Anonymization Benchmark provides a corpus of 1,268 court cases and metrics for privacy protection and utility preservation [14]. Our keep rate and over-removal measures follow the same concern for utility.

A third line of research measures risk through attacks. DIRI pairs a de-identification tool with a language model that attempts to match a redacted note to its patient. In a test of three tools on clinical notes from Weill Cornell Medicine, the attack still linked 9% of the notes from the best tool, ClinicalBERT, back to the correct patient [15]. RAT-Bench measures how many attributes an LLM attacker can infer from de-identified text and estimates re-identification risk across the US population. It finds that even the best tools are far from perfect [16]. Manzanares-Salor et al. train a machine learning attacker to evaluate the disclosure risk of de-identified documents [17]. These methods require an attacker model and measure risk after the fact. They do not provide a weighted score against a fixed answer key.

Real clinical text cannot be shared, so several studies use synthetic or partially synthetic text. Pissarra et al. added fake identifiers generated using the Faker library to MIMIC-III discharge summaries and compared Presidio, KNEO, a fine-tuned ClinicalBERT, and several Llama and Phi-2 models. ClinicalBERT performed best on their sensitivity metrics, and Presidio had the lowest result on four of the five [18]. RAT-Bench uses synthetic text based on US demographic figures, with direct and indirect identifiers across several domains and languages [16]. Neither study uses clinical trial narratives.

Our approach adds three things. Every item is assigned a rarity-based weight (S3) from the number of people who share its value. This is the same count that sits behind the 0.09 threshold, although the threshold is applied to combinations of attributes and S3 to one attribute at a time. The narratives are synthetic and come with an answer key generated from the records, so other groups can evaluate their own tools on the same type of data. And rankings are evaluated using a bootstrap, so a rank swap is called a confirmed reversal only when it persists after correction. We see the novelty in the first two together, not in either one alone. Weights based on counts can only be calculated exactly when the number of people sharing each value is known, and here it is, because the population is fictional and every item traces back to a row in it. The benchmark supplies the counts, and the weights give the benchmark a score that treats a missed gene variant differently from a missed mention of hypertension. The bootstrap is a standard check that keeps us from over-reading a swap, so we do not count it as a contribution. The weights allow a reader to see differences that F1 obscures, but our data do not demonstrate that they rank tools better.

Methodology

We evaluated four de-identification tools on synthetic clinical trial narratives. The question was whether the tools rank in a different order using flat F1 than using a risk-weighted F1 (RQ1). The hypothesis (H1) was that at least one pair of tools swaps positions between the two rankings and that the swap is statistically confirmed, meaning it is larger than chance (a confirmed reversal, defined under Statistical Analysis). The study consists of five components: a fictional population, trial records generated from it, narratives written from the records, an answer key with risk tiers, and an evaluation run of the four tools.

Plan Freeze

We froze the evaluation plan in a dated commit of the study repository (commit 065b623, tag plan-frozen) on October 21, 2026, prior to any tool running on the test set. The first tool output on the test set was written 23 minutes later. We did not register the plan with a third party. The answer key (commit dbcc418) and the tool settings (tag settings-frozen) had been frozen on October 12, 2026.

Synthetic Population

We generated a fictional population of 10,000 adults with a Python script (seed 20261005; Faker 40.39.0). Each person has a name, a date of birth, a region and a home town in one of eight countries (the United States, the United Kingdom, Germany, France, Spain, Poland, Canada, and Ireland), an occupation, and 0 to 3 of 17 common conditions. The conditions were drawn at published prevalence rates (Supplementary Table S1, Part A), and seven follow published age patterns (Supplementary Table S1, Part B) .Some people also received a rare condition, a relative's condition, an unusual procedure, a pregnancy outcome, a care-home stay, or a recent move from abroad. The 50 rare conditions were selected by a fixed rule from the Orphadata epidemiology file (June 2026 release) [19]: adult-onset diseases with a point prevalence between 1 to 9 per 1,000,000 and 1 to 5 per 10,000. Each was assigned to 1 to 5 people based on its prevalence rate. Three separate pools of 70 people received the

three trial diseases. For every value in the population we counted how many people share it (435 values). These counts establish the risk tiers and S3 weights described below.

Trial Records

We simulated three trials, one in oncology (A), one in a rare disease (B), and one in psychiatry (C). For each trial we selected 50 participants from its pool (seed 20261012) and generated one structured record per participant, following section 12.3.2 of the ICH E3 guideline [20]. A record contains subject and site identifiers, age and sex, the disease under study and its duration, medical history and concomitant medications, study drug and dose, one serious adverse event with its severity, clinical course, a key test result with its normal range, action taken, outcome, and causality. The study drugs (quelvatinib, pelvorisiran, and talmoraxetine) are invented, and an online search on October 6, 2026, found no real drug with these names. Each trial has 7 to 9 event types appropriate for the drug class and disease, and we verified event terms and grades against CTCAE v5 [21] and DSM-5 [22]. Every risk item is derived from the participant's own row in the population and contains the value it was derived from, so its count is known. Dates and study days follow from a first dose that falls within the participant's stated age (Day 1 is the first dose). Outcome, causality, and action taken use fixed wording. Five difficult cases (negation, second-hand claims, family history, abbreviations, and mixed date formats) were each assigned to 15 records, one difficult case per record. The 150 records contain 2,269 planted items, with 11 to 19 per record. Of these, 1,669 are risk items and 600 are Keep items, which are clinical content that must remain for the text to be useful.

Narratives

Claude Sonnet 5.5 generated one narrative per record through the claude.ai chat interface on October 11 and 12, 2026. Five prompt templates varied in style and length (one paragraph; two or three paragraphs; headed sections; formal report prose; clinical shorthand; 100 to 320 words). Each template was used for 30 records, assigned prior to generation, in its own chat outside any project, with five records per message. The generator received only the fields it required to generate. Country, person identifiers, and ISO dates were withheld. It enclosed every mention of a planted item in a tag containing the item's ID, so removing the tags yields the plain text and the character position of every mention. We refined the prompts over two pilot rounds of 15 narratives (one per template per trial). A script verified each round, and a separate chat reviewed it for clinical errors; the review led to corrections in the records. We never manually edited a narrative. Six narratives that failed a verification were regenerated. Appendix A illustrates one narrative with its items colored by risk tier.

Script Checks and Realism

A script that uses only regular expressions and table lookups, and none of the tools under evaluation, verified every narrative against its record. It verified that tags are well-formed and match the exact item text, that every planted item is tagged at every mention, and that no date, study day, age, ID-like number, title with a name, identifier from another record, or first name, surname, town, or region from the population is left untagged. It also verified that all ICH E3 elements are present with the record's exact outcome and causality wording, that the difficult case is visible, and that the timeline matches the record and is presented in order. On the first full run with the final prompts, 144 of 150 narratives passed every verification. After regeneration all 150 passed, and three length flags were cleared after reading because every item and element was present. Realism was verified by script only. We had planned a rating of 15 narratives by one experienced reviewer, but the reviewer did not respond, so we obtained no expert rating.

Answer Key and Risk Tiers

A script assigned each tagged item a risk tier (Table 1). Items that identify a single person on their own (subject and site IDs, investigator names) are tier 1. Calendar dates are tier 2 and study days are tier 3 by fixed rule. For every other item we counted how many of the 10,000 people share its value, which gives the item a count-based risk of 1 divided by that count. We used the 0.09 re-identification threshold of the European Medicines Agency and Health Canada [2] to set the first cut-off. A value shared by 11 people has a count-based risk of 1/11 = 0.0909, which is just above 0.09, and a value shared by 12 people has a risk of 0.083, which is below it. Tier 2 therefore contains values shared by 11 people or fewer, tier 3 values shared by 12 to 500 people, and tier 4 values shared by more than 500. The cut-offs of 12 to 500 and over 500 are our own choice, and we evaluated two others (11 and 200, and 11 and 1,000). Keep items carry no weight. The rules for tagging items and assigning tiers are in Appendix B. The author verified 50 randomly selected narratives (seed 20261020) against the draft key in Label Studio. None contained an untagged identifier (leak rate 0%) and no correction was required. We then froze the key (commit dbcc418) prior to any tool running.

Table 1. Items in the answer key by type and risk tier. The counts cover all 150 narratives, and the last column gives the 135 test narratives.

Item Type	Tier 1	Tier 2	Tier 3	Tier 4	Keep	All 150	Test 135
Common Condition	0	0	48	344	0	392	355
Calendar Date	0	260	0	0	0	260	233
Study Day	0	0	253	0	0	253	227
Subject ID	150	0	0	0	0	150	135
Age	0	1	149	0	0	150	135
Place	0	3	81	5	0	89	82
Occupation	0	8	34	39	0	81	72
Staff Name	74	0	0	0	0	74	67
Site ID	63	0	0	0	0	63	56
Relative's Condition	0	26	26	0	0	52	45
Gene Variant	0	12	25	0	0	37	34
Rare Diagnosis	0	33	0	0	0	33	30
Unusual Procedure	0	6	14	0	0	20	14
Recent Move	0	4	5	0	0	9	9
Care Home Stay	0	0	6	0	0	6	6
Keep Items	0	0	0	0	600	600	540
All Types	**287**	**353**	**641**	**388**	**600**	**2269**	**2040**

Tier 1: identifies a single person alone. Tier 2: shared by 11 people or fewer, or a calendar date. Tier 3: shared by 12 to 500 people, or a study day. Tier 4: shared by more than 500 people.

What the S3 Weights Measure

S3 is a rarity-based proxy. For an item whose value is shared by n people in the synthetic population, S3 is 1/n, capped at 1, and the same count decides the item's tier. It measures how unusual one value is. It has five limits.

First, it looks at one attribute at a time. The 0.09 threshold is meant for a group of people who share all the attributes an intruder knows. A narrative carries 11 to 19 items, and the people who match all of them cannot be more numerous than the people who match the rarest one, and are usually far fewer. For an intruder who knows every item, the risk for the whole narrative is therefore at least the S3 weight of its rarest item and is usually higher. S3 does not say how much higher.

Second, the weights are not probabilities that can be added. RWR is the share of the total weight that a tool removed. It shows how much of the rare material was caught and does not give the chance that a person is re-identified.

Third, the counts come from a fictional population of 10,000 people built by our own rules. Real values would have different counts, and the population reproduces real-world patterns only where Supplementary Table S1 says so.

Fourth, S3 does not model what a particular intruder knows or can look up, which is the idea behind the intruder types described by Graff [11]. Fifth, dates and study days have no count and take the median weight of their tier (see Evaluation).

For these reasons we use S3 to compare how the tools treat rare and common items. We do not use it to say that a narrative is safe or unsafe to release.

De-Identification Tools

Four tools ran once on every test narrative. We selected their settings on 15 setup narratives (5 per trial, seed 20261019), kept the other 135 narratives (45 per trial) for testing, and froze the settings prior to the test run. Presidio 2.2.364 [23] with the spaCy model en_core_web_lg 3.8.0 used the entity types DATE_TIME, LOCATION, NRP, PERSON, and PHONE_NUMBER at a score threshold of 0.2. Both choices came from a rule established in advance: keep the types for which at least half the removed characters lie in risk items, then select the threshold with the highest character F1. We included the scispaCy model en_ner_bc5cdr_md 0.5.4 (scispaCy 0.6.2 [24]) as a disease and chemical named-entity recognition (NER) baseline. It was trained on the BioCreative V CDR corpus [25] and is not a de-identification tool. We call it the scispaCy NER baseline, and scispaCy NER in the tables. We counted every DISEASE and CHEMICAL entity it found as removed. The RoBERTa de-identification model obi/deid_roberta_i2b2 [26] (revision 78f2152), called RoBERTa in the tables, was trained on US hospital notes, and it read chunks of whole sentences of at most 500 tokens. The language model qwen3:8b (LLM in the tables) ran locally through Ollama 0.40.6 with temperature 0, seed 20261026, JSON output, and a context of 4,096 tokens. Its standard prompt asked for the exact strings to remove (Appendix C). We mapped each returned string to every exact, case-sensitive occurrence in the narrative and counted strings that do not occur in the text as invalid. For every tool we merged overlapping and touching spans, so each output is a list of character spans.

Evaluation

A script converted each output into a set of removed characters and evaluated it against the frozen key, ignoring whitespace in every count. Under the strict rule, an item counts as caught only when every non-whitespace character of every mention is removed, because one leaked mention can identify a person [14]. A loose rule, which counts any overlap, served as a sensitivity check. Precision (P) is counted at the character level: it is the share of removed characters that fall inside a tier 1 to 4 item. Flat recall is counted at the item level: it is the share of risk items caught under the matching rule, and every risk item has the same weight. Flat F1 is the harmonic mean of this character-level precision and this item-level recall. It differs from the conventional span-level F1, where precision and recall are both computed over matched spans, so our flat F1 should not be compared directly with span-level F1 values from other studies. For risk-weighted recall (RWR), item i has weight w_i, and c_i is 1 if the item was caught and 0 if not:

$$RWR = \Sigma\, w_i c_i \,/\, \Sigma\, w_i \quad (1)$$

$$RW\text{-}F1 = 2 \cdot P \cdot RWR \,/\, (P + RWR) \quad (2)$$

We used three weighting schemes. S1 gives every risk item weight 1, so RWR equals flat recall and RW-F1 under S1 equals flat F1. S2 gives tiers 1 to 4 the points 5, 4, 2, and 1. S3 gives each item 1 divided by the number of people in the synthetic population who share its value, capped at 1. It is a rarity proxy built from the frequency of single attributes, and the subsection What the S3 Weights Measure sets out what it can and cannot say about risk. S3 is the main scheme. Dates and study days have no population count, so they take the median S3 weight of the counted items in their tier (0.5 and 0.006757). The scheme builds on the risk-based evaluation of Scaiano et al. [9] and extends it from identifier types to graded weights on every item, including medical details. Next to the privacy scores we report two measures of usefulness. The keep rate is the share of Keep items left untouched (higher is better). Over-removal is the number of removed characters that belong to no risk item, as a share of all characters (lower is better). Classic PHI F1 uses only the item types subject ID, site ID, staff name, calendar date, age, and place, which is how tools are usually reported. All scores are pooled over every item or character in the 135 test narratives.

Statistical Analysis

We ranked the four tools by flat F1 and by RW-F1 under S3 (strict matching, tier cut-offs of 11 and 500 people) and compared the two rankings with Kendall's tau-b. To see how much the scores depend on the set of narratives, we resampled narratives with replacement 2,000 times within each trial (45 per trial each time; seed 20261102) and recomputed every score from the pooled counts [27]. For each of the six tool pairs we took the difference in flat F1 (ΔF) and in S3 RW-F1 (ΔR), with a 95% percentile interval and a two-sided bootstrap p-value. We adjusted the p-values with the Holm correction [28], separately for the six flat-F1 tests and the six RW-F1 tests. A pair counted as a confirmed reversal when both adjusted p-values were below

0.05 and ΔF and ΔR had opposite signs. A rank swap that does not meet this rule is reported as an observed swap and not as a reversal. H1 was supported if at least one pair showed a confirmed reversal. For each pair we also report the swap share, which is the share of resamples in which ΔF and ΔR had opposite signs. Sensitivity checks repeated the analysis with S2 in place of S3, with tier cut-offs of 11 and 200 and of 11 and 1,000 (each under S3 and S2), and with loose matching under S3, all on the same 2,000 resamples.

Results by Trial, Catch Rates by Item Type and by Tier, and Stability of LLM's Output

Results by trial, catch rates by item type and by tier, and stability of the LLM's output are exploratory. We ran the LLM a second time with the same settings on 25 randomly selected test narratives (seed 20261021). For these we report the share of risk items with the same strict caught status in both runs, Cohen's kappa [29], and overlap of removed characters, which is the number of characters removed in both runs divided by the number removed in either run.

Computing Environment

All four tools were run on a single laptop with an Intel Core i5-10210U processor (1.6 GHz base clock), 16 GB of memory and integrated Intel UHD graphics, running a 64-bit Windows system. It took Presidio 11 seconds, the scispaCy NER baseline 9 seconds and the RoBERTa model 639 seconds to process the 135 test narratives. The LLM took 8,101 seconds (2.25 hours), or an average of 60 seconds per narrative. Analysis was performed using Python 3.12.10 with numpy 2.5.4, pandas 3.0.6 and scipy 1.18.1; matplotlib 3.11.2 was used to create the figure.

Deviations From the Plan

The study deviated from the frozen plan in several ways. After the first pilot, the generator added statements to the narratives indicating that information was not reported, so we reduced the word limits to what could be filled by the records (templates 1 to 5: 150 to 220, 200 to 280, 200 to 300, 220 to 320 and 100 to 180 words). Only four participants in each of trials A and C had a relative with a rare disease, so the family-history difficult case was split 4, 7 and 4 across the trials instead of 5, 5 and 5. The population allows each person at most three common conditions, so narratives contain two to three of them (the plan stated usually four), plus one negated or second-hand condition in 30 records. One full generation run on 12 October 2026 used incorrect prompt versions for four templates, and we did not use it. Pilot reviews resulted in corrections to the records, and all 150 narratives were generated from the final records. We used seed 20261021 instead of the planned 20261027 to select the 25 narratives for the LLM re-run, and we made this change before the sample was drawn. Although the frozen plan specifies the population and record seeds as 20261004 and 20261006, the scripts and the run log use 20261005 and 20261012, and the data were created with those two. We did not update the frozen plan. The RoBERTa model processed chunks of at most 500 tokens instead of 512, to allow room for the model's two special tokens. There was no public registration and no expert realism rating, as described above.

Results

Dataset & Checks

The 150 narratives contain 2,269 tagged items: 1,669 risk items and 600 Keep items (Table 1). The 135 test narratives contain 2,040 items, of which 1,500 are risk items and 540 are Keep items. All 150 narratives passed every script check. The hand check of 50 narratives revealed no untagged identifier (0 of 50). Since no expert evaluated realism, the realism check is based solely on the script checks.

Main Ranking

Table 2 and Figure 1 display the scores of the four tools on the 135 test narratives. Under both flat F1 and S3 RW-F1, the LLM ranked first and the scispaCy NER baseline ranked last. The two tools in the middle swapped positions in the point estimates. The RoBERTa model ranked higher than Presidio on flat F1 (0.465 against 0.462), while Presidio ranked higher on RW-F1 (0.495 against 0.467). This is an observed rank swap. The next subsection shows that it is not a statistically confirmed reversal. Kendall's tau-b between the two rankings is 0.667, as one of the six pairs is in the opposite order in the two rankings.

The remaining columns of Table 2 illustrate how the tools differ. The RoBERTa model had the highest precision and the highest keep rate, but it removed few risk items. The LLM's score increases from flat F1 to RW-F1, the scispaCy NER baseline's decreases, and the RoBERTa model's remains relatively constant. The scispaCy NER baseline had the lowest precision, the lowest keep rate and the greatest over-removal, and its RW-F1 is near zero. Classic PHI F1 ranks the RoBERTa model first, the LLM second and Presidio third, therefore a tool selected based on the traditional identifier score would not be the same as either F1.

Table 2. Scores of the four tools on the 135 test narratives under strict matching. Flat F1 and RW-F1 show 95% bootstrap intervals. Flat F1 combines character-level precision with item-level recall. RWR is risk-weighted recall. RW-F1 and RWR use the S3 weights. RoBERTa is the RoBERTa de-identification model obi/deid_roberta_i2b2, and scispaCy NER is the scispaCy disease and chemical NER baseline.

Tool	Precision	Flat recall	Flat F1	RWR (S3)	RW-F1 (S3)	Keep rate	Over-removal	Classic PHI F1
LLM	0.802	0.539	0.645 [0.619, 0.670]	0.824	0.813 [0.770, 0.854]	0.943	0.016	0.723
RoBERTa	0.930	0.310	0.465 [0.451, 0.479]	0.312	0.467 [0.438, 0.495]	0.998	0.004	0.757
Presidio	0.723	0.339	0.462 [0.446, 0.478]	0.376	0.495 [0.466, 0.522]	0.993	0.018	0.663
scispaCy NER	0.306	0.227	0.261 [0.247, 0.274]	0.034	0.062 [0.037, 0.087]	0.444	0.109	0.006

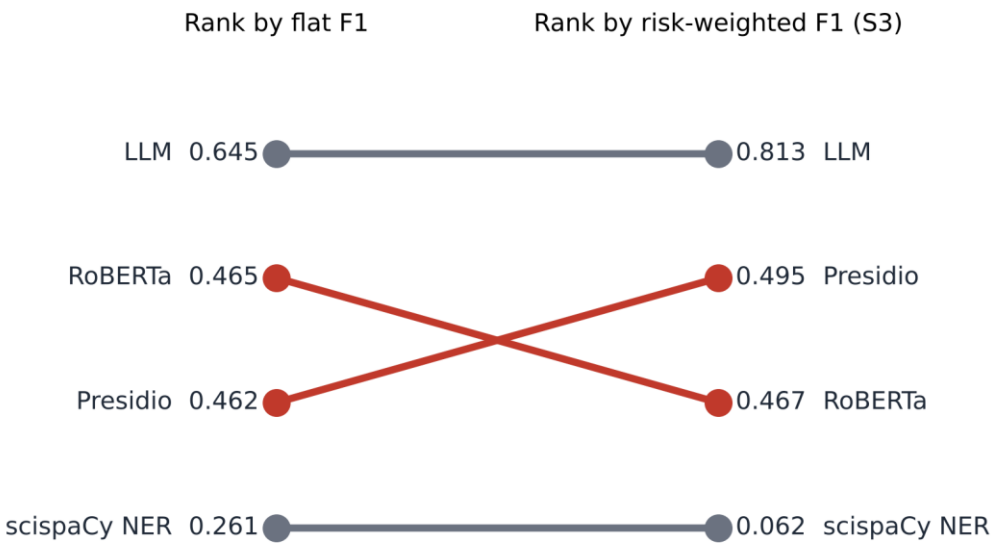

Figure 1. Rank change of the four tools between flat F1 (left) and S3 RW-F1 (right) on the 135 test narratives. Each line joins a tool's rank on the two scores. The crossing red lines show the observed Presidio and RoBERTa swap, which is not a statistically confirmed reversal (Table 3).

Test of H1

H1 was not supported. Only Presidio and the RoBERTa model swapped positions. This is an observed rank swap and not a statistically confirmed reversal, because the bootstrap cannot distinguish it from chance (Table 3). The confidence intervals of both differences include zero, and the Holm-adjusted p-values are 0.709 for flat F1 and 0.208 for RW-F1. ΔF and ΔR had opposite signs in 54% of the resamples. The other five pairs never swapped, and every difference between them was statistically significant after correction (adjusted p = 0.006 for both scores, the smallest possible value for 2,000 resamples). There was no confirmed reversal.

Table 3. Differences between tool pairs under strict matching. Each difference is the first tool minus the second, with

a 95% bootstrap interval. The swap share is the share of 2,000 resamples in which ΔF and ΔR had opposite signs.

Tool pair	ΔF, flat F1 (95% interval)	ΔR, S3 RW-F1 (95% interval)	Swap share
Presidio vs scispaCy NER	0.201 [0.180, 0.224]	0.433 [0.394, 0.471]	0%
Presidio vs RoBERTa	−0.003 [−0.019, 0.013]	0.028 [−0.013, 0.070]	54%
Presidio vs LLM	−0.183 [−0.210, −0.154]	−0.318 [−0.367, −0.268]	0%
scispaCy NER vs RoBERTa	−0.204 [−0.226, −0.183]	−0.405 [−0.446, −0.364]	0%
scispaCy NER vs LLM	−0.384 [−0.415, −0.354]	−0.752 [−0.797, −0.700]	0%
RoBERTa vs LLM	−0.180 [−0.209, −0.149]	−0.346 [−0.398, −0.291]	0%

Stability Checks

No row in Table 4 has a confirmed reversal, so H1 was not supported in any of the seven scenarios. At all three cut-offs (tau-b of 1.000), the tools maintain the flat-F1 order under S2, and the maximum swap share is 10% to 11%. Under S3 with cut-offs of 11 and 200 and of 11 and 1,000, the observed Presidio and RoBERTa swap persists, with a swap share of 54% in both. The loose-matching row is different. The swapped pair there is the RoBERTa model and the LLM, with a swap share of 94%. In that row, the LLM leads on flat F1 by 0.041 (adjusted p = 0.008) and the RoBERTa model leads on RW-F1 by 0.037 (adjusted p = 0.106). The first difference is significant after correction and the second is not, therefore we report the loose-matching result as a trend and not as a confirmed reversal.

Table 4. Stability of the ranking under the sensitivity checks. The swap share is the highest among the six tool pairs.

Check	Kendall tau-b	Observed swap	Highest swap share	Confirmed reversal	H1 supported
Main: S3, strict, cut-offs 11/500	0.667	Presidio vs RoBERTa	54%	None	No
S2 instead of S3	1.000	None	10%	None	No
Cut-offs 11/200, S3	0.667	Presidio vs RoBERTa	54%	None	No
Cut-offs 11/200, S2	1.000	None	11%	None	No
Cut-offs 11/1,000, S3	0.667	Presidio vs RoBERTa	54%	None	No
Cut-offs 11/1,000, S2	1.000	None	10%	None	No
Loose matching, S3	0.667	RoBERTa vs LLM	94%	None	No

Exploratory Results

Results by trial. The LLM ranked first and the scispaCy NER baseline last in each trial under both scores (Table 5). The order of Presidio and the RoBERTa model depended on the trial. In trials A and C, Presidio ranked above the RoBERTa model under both scores. In trial B, the RoBERTa model ranked above Presidio on flat F1 and below it on RW-F1, therefore the observed swap appears only in trial B.

Table 5. Scores and ranks of the four tools in each trial (45 test narratives per trial, strict matching).

Trial	Tool	Flat F1	Rank by flat F1	RW-F1 (S3)	Rank by RW-F1
A	LLM	0.642	1	0.777	1
A	Presidio	0.489	2	0.522	2
A	RoBERTa	0.484	3	0.476	3
A	scispaCy NER	0.260	4	0.063	4
B	LLM	0.641	1	0.827	1
B	RoBERTa	0.419	2	0.399	3
B	Presidio	0.403	3	0.440	2
B	scispaCy NER	0.267	4	0.069	4
C	LLM	0.655	1	0.843	1
C	Presidio	0.506	2	0.531	2
C	RoBERTa	0.496	3	0.524	3
C	scispaCy NER	0.250	4	0.052	4

Catch rates by item type. Table 6 provides the strict catch rate of each tool for every item type in the test set. The scispaCy NER baseline identified most common conditions (0.896) and slightly more than half of the rare diagnoses (0.567), and virtually nothing else. The LLM identified nearly all subject IDs, site IDs and calendar dates. The RoBERTa model and Presidio remove the number from a subject or site ID but retain the label word, therefore their strict catch rates for subject IDs are 0.141 and 0, compared to 1.000 and 0.919 under loose matching. No tool identified more than 0.088 of occupations, unusual procedures, relatives' conditions or gene variants. The numbers of care home stays (6), recent moves (9) and procedures (14) are too small to interpret.

Table 6. Strict catch rate by item type on the 135 test narratives. Items is the number of risk items of that type.

Item type	Items	RoBERTa	LLM	Presidio	scispaCy NER
Common condition	355	0.000	0.054	0.006	0.896
Calendar date	233	0.940	0.991	0.807	0.004
Study day	227	0.000	0.714	0.163	0.000
Subject ID	135	0.141	1.000	0.000	0.000
Age	135	1.000	0.533	1.000	0.015
Place	82	1.000	0.939	0.927	0.037
Occupation	72	0.000	0.042	0.000	0.000
Staff name	67	0.000	0.597	0.925	0.000

Item type	Items	RoBERTa	LLM	Presidio	scispaCy NER
Site ID	56	0.000	0.982	0.000	0.000
Relative's condition	45	0.000	0.067	0.000	0.000
Gene variant	34	0.029	0.088	0.000	0.000
Rare diagnosis	30	0.000	0.033	0.000	0.567
Unusual procedure	14	0.000	0.000	0.000	0.000
Recent move	9	1.000	0.556	1.000	0.000
Care home stay	6	0.000	0.500	0.000	0.000

Stability of the LLM. On the 25 re-run narratives, the two runs of the LLM produced identical strict results for 276 of the 280 risk items (98.6%; Cohen's kappa 0.971), and the agreement under loose matching was the same. Of the 25 outputs, 19 were identical, and the average overlap of removed characters was 0.96. Of the 902 strings the LLM returned on the test set, 4 (0.44%) did not appear in the narrative. No narrative produced unreadable JSON.

Discussion

The selection of score affected how far apart the tools appeared more than it affected their order. The LLM led under both scores, and its lead over the next tool increased from 0.180 on flat F1 to 0.318 on RW-F1 (Table 2). The scispaCy NER baseline decreased from 0.261 to 0.062, while the RoBERTa model remained relatively unchanged. A sponsor selecting a tool based on classic PHI F1 would have selected the RoBERTa model, which neither F1 selects (Table 2). Reporting only one F1 would obscure this.

H1 was not supported for a simple reason. Presidio and the RoBERTa model are within 0.003 of each other on flat F1, and the RW-F1 difference of 0.028 has a confidence interval that includes zero. The observed swap between them is a difference in point estimates and not a confirmed reversal. With 135 narratives we could only confirm a reversal between tools whose scores are substantially different, and none of the far-apart pairs changed order. The swap appears only in trial B (Table 5), so it may depend on the type of narrative. A larger study with 450 narratives would narrow the confidence intervals, and whether Presidio and the RoBERTa model would then separate is an open question.

The loose-matching row is a trend and requires caution. In that row, the LLM leads on flat F1 by 0.041 (adjusted $p = 0.008$) and the RoBERTa model leads on RW-F1 by 0.037 (adjusted $p = 0.106$), therefore only the first difference is significant after correction. Loose matching also awards credit for removing the number in an ID while retaining its label, which is a rule choice and could account for some of the RoBERTa model's improvement.

The most useful result for a practitioner is presented in Table 6. Occupations, unusual procedures, relatives' conditions and gene variants are details that can make an individual rare, and no tool identified more than 0.088 of them. All relatives' conditions, gene variants and unusual procedures fall into tiers 2 and 3 (Table 1), where the weights are greatest. Rare diagnoses were identified only by the scispaCy NER baseline (0.567), a tagger that identifies diseases in general and that also removed more than half of the Keep items. A tool can perform well on classic identifiers and still leave these details in the text.

The label words also matter. The RoBERTa model and Presidio remove the number from a subject or site ID and retain the word Subject or Site, therefore their strict catch rates for subject IDs are 0.141 and 0, compared to 1.000 and 0.919 under loose matching. Which rule is appropriate depends on the sponsor's policy. We report both.

Our results are consistent with prior research without contradicting it. Pissarra et al. found ClinicalBERT to be superior to Presidio on most sensitivity metrics in MIMIC-III notes [18]. We found the RoBERTa de-identification model ahead of Presidio on flat F1 by a margin within chance, and Presidio to be superior on RW-F1. The data, tools and scores differ, therefore the two results indicate primarily that a ranking is dependent on the notes and on the score. Graff et al. reported that psychiatry is difficult for dictionary-based NLP systems [10]. We observed no indication that trial C, our psychiatry trial, was more difficult for the tools we evaluated (Table 5). Our tools are not dictionary-based systems and our narratives are synthetic, therefore this is not an evaluation of their finding. Aghakasiri et al. cautioned that F1 conceals clinical information loss [6]. Our keep rate

demonstrates the same phenomenon: the scispaCy NER baseline retained only 44% of Keep items and had the highest over-removal of the four tools (Table 2). Attack-based studies such as DIRI and RAT-Bench [15], [16] assess what an attacker can recover, which our scores do not. An attacker assessment is planned as a subsequent study.

In reality, the sponsor faces a trade-off among three parameters: the residual risk of the text, the amount of clinical content remaining after processing, and the computational and financial costs related to the use of the tool. The research provides results concerning each of them individually (Tables 2 and 6). The RoBERTa model preserved 99.8% of Keep items while throwing away nearly no item, which was not an identifier, however, only approximately one-third of the risk weight (RWR 0.312). The LLM managed to capture the majority of the risk (RWR 0.824), however, at the cost of deleting 5.7% of Keep items as well as requiring 8,101 seconds of processing 135 narratives on a laptop. The Presidio took 11 seconds and preserved 99.3% of Keep items while having RWR equal to 0.376. No tool performed perfectly well concerning all the three parameters mentioned above; therefore, the selection will depend on the acceptable residual risk and the time and computational power. Since the EMA requires the preservation of adverse event terms [1], the keep rate should be used together with the privacy score.

The residual risk deserves the most attention because even the LLM leaves approximately 18% of the risk weight in the text, and the tools have captured practically none of the rare items mentioned above (Table 6). The LLM had different outputs for two runs for six out of 25 narratives, and the variability will need to be checked by sponsors before using the tool for one run. All the results are based on the synthetic narratives, one small model, and one laptop; therefore, the sponsors are recommended to replicate the comparison using their own narratives before applying it. The answer key and scoring methodology are aimed at the ease of such replication. In principle, the population-derived weights allow translating the differential weighting into the testable formula. The next logical step is applying weights by intruder type.

Conclusion

We evaluated four de-identification tools on 135 synthetic clinical trial narratives using flat F1 and a risk-weighted F1 whose weights are based on counts in a hypothetical population of 10,000 individuals. The LLM achieved the highest score under both scores. Presidio and the RoBERTa model swapped positions in the point estimates, but this observed swap is not a statistically confirmed reversal, and H1 was not supported in any of the seven scenarios. The weights did alter the magnitude of differences between tools, and they demonstrated that all four tools fail to identify more than 90% of occupations, unusual procedures, relatives' conditions and gene variants.

The study has limitations. One model authored each narrative, therefore the style may be more uniform than in actual narratives. Rare diagnoses originate from the synthetic population and appear without their typical treatments or associated findings. No clinical expert reviewed the narratives, therefore realism is assessed solely through script checks. There was no second annotator: scripts reviewed all 150 narratives and the author reviewed 50. The plan was frozen in a dated commit and was not registered with a third party. We evaluated three therapeutic areas, four tools and one small LLM, and model versions evolve, therefore precise scores will become dated. The 135 test narratives provide wide confidence intervals. The LLM produced a different output on a second run for 6 of 25 narratives, a variation that the bootstrap does not capture. Strict matching awards credit for label words such as Subject and Dr, therefore a tool that removes only the number receives partial credit. The RoBERTa model was trained on US hospital notes, the scispaCy NER baseline is a disease and chemical tagger and not a de-identification tool, and date style is linked to the narrative template. The S3 weights are rarity proxies for single attributes and do not estimate the re-identification risk of a whole narrative. Our flat F1 combines character-level precision with item-level recall, so it cannot be compared directly with span-level F1 from other studies.

The next steps include a larger study with 450 narratives, which would decrease the confidence intervals and potentially resolve the Presidio and RoBERTa pair, an attacker assessment in which a language model attempts to re-identify individuals from each tool's output (RQ2), and a risk-aware prompt for the LLM that incorporates the weights (RQ4).

In addition to this study, the methodology provides sponsors and regulators with a means to inquire what items a tool fails to identify, and not merely how many. Because the narratives are synthetic, other organizations can develop similar types of data sets and evaluate their own tools without accessing patient data.

Declarations

Ethics. All people in the data are fictional, and no real patient data were used. No ethics review was requested.

Data and code availability. The materials needed to reproduce Table 2 and Table 3 are provided as Supplementary Material S2, a reviewer package submitted with this manuscript (risk_weighted_deid_reviewer_package.zip). It contains the 135 frozen

test narratives as plain text, the frozen answer key and the setup and test split, the tool settings, the output of each tool on the test narratives, the scoring and statistics scripts, the analysis outputs behind Table 2 and Table 3, the environment files with exact package versions, and a README with the commands that reproduce the scores. The package matches the frozen repository state: tag plan-frozen (commit 065b623), tag settings-frozen and answer key commit dbcc418. All people and narratives are synthetic, so the package holds no patient data.

Use of AI. Claude Sonnet 5.5 (Anthropic) wrote the synthetic narratives, as described in the Methodology, and qwen3:8b was one of the four tools tested. The author also used Claude to help review the literature, to research related work, to debug code, to draft parts of the text and to check grammar and wording.

References

1. European Medicines Agency, "External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use," EMA/90915/2016, 2016.

2. Health Canada, "Guidance document on public release of clinical information," Ottawa, ON, Canada, 2019.

3. Ö. Uzuner, Y. Luo, and P. Szolovits, "Evaluating the state-of-the-art in automatic de-identification," J. Am. Med. Inform. Assoc., vol. 14, no. 5, pp. 550–563, 2007.

4. A. Stubbs, C. Kotfila, and Ö. Uzuner, "Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1," J. Biomed. Inform., vol. 58, pp. S11–S19, 2015.

5. M. Sariyar, "Clinical text de-identification beyond PHI detection: A scoping review," J, vol. 9, no. 3, Art. no. 22, 2026, doi: 10.3390/j9030022.

6. K. Aghakasiri et al., "Not what the doctor ordered: Surveying LLM-based de-identification and quantifying clinical information loss," in Proc. 2025 Conf. Empirical Methods in Natural Language Processing (EMNLP), 2025, pp. 32187–32203.

7. G. Loukides, J. C. Denny, and B. Malin, "The disclosure of diagnosis codes can breach research participants' privacy," J. Am. Med. Inform. Assoc., vol. 17, no. 3, pp. 322–327, 2010.

8. J. Branson, N. Good, J.-W. Chen, W. Monge, C. Probst, and K. El Emam, "Evaluating the re-identification risk of a clinical study report anonymized under EMA Policy 0070 and Health Canada Regulations," Trials, vol. 21, no. 1, Art. no. 200, Feb. 2020, doi: 10.1186/s13063-020-4120-y.

9. M. Scaiano, G. Middleton, L. Arbuckle, V. Kolhatkar, L. Peyton, M. Dowling, D. S. Gipson, and K. El Emam, "A unified framework for evaluating the risk of re-identification of text de-identification tools," J. Biomed. Inform., vol. 63, pp. 174–183, 2016, doi: 10.1016/j.jbi.2016.07.015.

10. N. Graff, G. Nenadic, M. Elliot, and A. M. Hughes, "Beyond standard metrics: Assessing NLP for anonymisation of clinical trial narratives," presented at the UNECE Expert Meeting on Statistical Data Confidentiality, Oct. 2025.

11. N. M. Graff, "Data privacy and anonymisation in clinical trial data sharing," Ph.D. dissertation, Univ. of Manchester, Manchester, U.K., 2025.

12. S. M. Meystre, F. J. Friedlin, B. R. South, S. Shen, and M. H. Samore, "Automatic de-identification of textual documents in the electronic health record: A review of recent research," BMC Med. Res. Methodol., vol. 10, Art. no. 70, 2010.

13. O. Ferrández, B. R. South, S. Shen, F. J. Friedlin, M. H. Samore, and S. M. Meystre, "Evaluating current automatic de-identification methods with Veteran's health administration clinical documents," BMC Med. Res. Methodol., vol. 12, Art. no. 109, 2012.

14. I. Pilán, P. Lison, L. Øvrelid, A. Papadopoulou, D. Sánchez, and M. Batet, "The Text Anonymization Benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization," Comput. Linguist., vol. 48, no. 4, pp. 1053–1101, 2022, doi: 10.1162/coli_a_00458.

15. J. X. Morris et al., "DIRI: Adversarial patient reidentification with large language models for evaluating clinical text anonymization," AMIA Jt. Summits Transl. Sci. Proc., vol. 2025, pp. 355–364, 2025.

16. N. Krčo, Z. Yao, M. Meeus, and Y.-A. de Montjoye, "RAT-Bench: A comprehensive benchmark for text anonymization," arXiv:2602.12806, 2026.

17. B. Manzanares-Salor, D. Sánchez, and P. Lison, "Evaluating the disclosure risk of anonymized documents via a machine learning-based re-identification attack," Data Min. Knowl. Discov., vol. 38, pp. 2559–2594, 2024.

18. D. Pissarra, I. Curioso, J. Alveira, D. Pereira, B. Ribeiro, T. Souper, V. Gomes, A. V. Carreiro, and V. Rolla, "Unlocking the potential of large language models for clinical text anonymization: A comparative study," arXiv:2406.00062, 2024.

19. Orphanet, "Orphadata: Epidemiological data on rare diseases," June 2026 release. [Online]. Available: https://www.orphadata.com

20. International Council for Harmonisation, "Structure and content of clinical study reports E3," ICH Harmonised Tripartite Guideline, 1995.

21. National Cancer Institute, "Common Terminology Criteria for Adverse Events (CTCAE), version 5.0," U.S. Dept. of Health and Human Services, Nov. 2017.

22. American Psychiatric Association, Diagnostic and Statistical Manual of Mental Disorders, 5th ed. Arlington, VA, USA: American Psychiatric Publishing, 2013.

23. Microsoft, "Presidio: Data protection and de-identification SDK." [Online]. Available: https://github.com/microsoft/presidio

24. M. Neumann, D. King, I. Beltagy, and W. Ammar, "ScispaCy: Fast and robust models for biomedical natural language processing," in Proc. 18th BioNLP Workshop and Shared Task, Florence, Italy, 2019, pp. 319–327.

25. J. Li et al., "BioCreative V CDR task corpus: A resource for chemical disease relation extraction," Database, vol. 2016, Art. no. baw068, 2016.

26. obi, "deid_roberta_i2b2," Hugging Face model card. [Online]. Available: https://huggingface.co/obi/deid_roberta_i2b2

27. B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap. New York, NY, USA: Chapman & Hall, 1993.

28. S. Holm, "A simple sequentially rejective multiple test procedure," Scand. J. Stat., vol. 6, no. 2, pp. 65–70, 1979.

29. J. Cohen, "A coefficient of agreement for nominal scales," Educ. Psychol. Meas., vol. 20, no. 1, pp. 37–46, 1960.

Appendices

Supplementary Table S1

Supplementary Table S1. Part A. The 17 common conditions in the fictional population of 10,000 people. Each condition was drawn at the target figure for every adult alike, which is a simplification because most figures refer to a specific age range. The cap of 3 conditions per person and the rule that trial-pool members have at least 2 move the achieved counts slightly away from the target. Tier is the risk tier under the main cut-offs of 11 and 500 people.

Condition	Target	Achieved count	Achieved %	Tier	Source
hypertension	33.00%	3,084	30.84%	4	WHO. Global report on hypertension: the race against a silent killer. Geneva: WHO; 2023. 33% of adults aged 30-79.

Condition	Target	Achieved count	Achieved %	Tier	Source
tension-type headache	26.00%	2,527	25.27%	4	Stovner LJ et al. The global prevalence of headache: an update. J Headache Pain 2022;23:34. Migraine 14.0%, tension-type headache 26.0%.
obesity	23.00%	2,166	21.66%	4	WHO Regional Office for Europe. WHO European Regional Obesity Report 2022. 23% of adults.
migraine	14.00%	1,404	14.04%	4	Stovner LJ et al. The global prevalence of headache: an update. J Headache Pain 2022;23:34. Migraine 14.0%, tension-type headache 26.0%.
gastro-oesophageal reflux disease	13.30%	1,285	12.85%	4	Eusebi LH et al. Global prevalence of, and risk factors for, gastro-oesophageal reflux symptoms: a meta-analysis. Gut 2018;67:430-440. 13.3% (weekly symptoms).
chronic obstructive pulmonary disease	10.30%	894	8.94%	4	Adeloye D et al. Global, regional, and national prevalence of COPD in 2019. Lancet Respir Med 2022;10:447-458. 10.3% of people aged 30-79 (GOLD definition).
chronic kidney disease	10.00%	911	9.11%	4	Kovesdy CP. Epidemiology of chronic kidney disease: an update 2022. Kidney Int Suppl 2022;12:7-11. More than 10% of the general population.
type 2 diabetes mellitus	10.00%	935	9.35%	4	International Diabetes Federation. IDF Diabetes Atlas, 11th edition. Brussels: IDF; 2025. 11.1% of adults aged 20-79 had diabetes in 2024, and over 90% of people with diabetes have type 2, so 10.0% is used.
osteoarthritis	7.60%	676	6.76%	4	GBD 2021 Osteoarthritis Collaborators. Lancet Rheumatol 2023;5:e508-e522. 7.6% of the global population in 2020.
anxiety disorder	5.80%	551	5.51%	4	WHO. Anxiety disorders, fact sheet, 15 September 2026 (GBD 2023). 5.8% of the global population.
depression	5.20%	490	4.90%	3	WHO. Depressive disorder (depression), fact sheet, 11 September 2026 (GBD 2023). 5.2% of adults.

Condition	Target	Achieved count	Achieved %	Tier	Source
asthma	3.40%	326	3.26%	3	Global Asthma Network. The Global Asthma Report 2022 (GBD 2019). 3,416 per 100,000 (age-standardised).
atrial fibrillation	3.00%	277	2.77%	3	Hindricks G et al. 2020 ESC Guidelines for the diagnosis and management of atrial fibrillation. Eur Heart J 2021;42:373-498. 2-4% of adults; midpoint used.
heart failure	1.50%	132	1.32%	3	McDonagh TA et al. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. Eur Heart J 2021;42:3599-3726. 1-2% of adults; midpoint used.
coeliac disease	0.70%	57	0.57%	3	Singh P et al. Global prevalence of celiac disease: systematic review and meta-analysis. Clin Gastroenterol Hepatol 2018;16:823-836. 0.7% (biopsy-confirmed).
epilepsy	0.64%	49	0.49%	3	Fiest KM et al. Prevalence and incidence of epilepsy. Neurology 2017;88:296-303. Active epilepsy 6.38 per 1,000.
rheumatoid arthritis	0.46%	44	0.44%	3	Almutairi K et al. The global prevalence of rheumatoid arthritis: a meta-analysis. Rheumatol Int 2021;41:863-877. 460 per 100,000.

Supplementary Table S1. Part B. The seven conditions that follow a published age pattern. The pattern sets how much more likely an older person is to have the condition, and the chances are then scaled so that the share across all adults matches the target in Part A. The age-group columns show the achieved share in this population. Where the source column says a value is an assumption, the published figures did not cover that age group.

Condition	Published pattern	18-29	30-39	40-49	50-59	60-69	70-79	80-100	Source
hypertension	18-39: 22.4%; 40-59: 54.5%; 60-100: 74.5%	14.8%	14.9%	34.2%	33.2%	45.5%	43.4%	44.6%	Ostchega Y et al. Hypertension prevalence among adults aged 18 and over: United States, 2017-2018. NCHS Data Brief No. 364, 2020. 22.4% at 18-39, 54.5% at 40-59, 74.5% at 60+.
type 2 diabetes mellitus	18-24: 0.2%; 25-34: 1.3%; 35-49: 5.3%; 50-64: 13.3%; 65-100: 20.1%	0.5%	3.9%	5.6%	13.1%	14.6%	17.5%	20.5%	CDC. QuickStats: Percentage of adults aged 18+ with diagnosed diabetes, by urbanization level and age group, NHIS 2022. MMWR 2024;73(2). 1.3% at 18-34, 5.3% at 35-49 and 13.3% at 50-64 (metropolitan), 20.1% at 65+.

Condition	Published pattern	18-29	30-39	40-49	50-59	60-69	70-79	80-100	Source
chronic obstructive pulmonary disease	18-24: 0.0%; 25-29: 1.0%; 30-39: 3.8%; 40-49: 6.3%; 50-59: 10.2%; 60-69: 15.8%; 70-100: 23.4%	0.4%	3.4%	6.6%	9.8%	13.6%	20.0%	19.6%	Adeloye D et al. Lancet Respir Med 2022;10:447-458, Table 1 (GOLD-COPD, worldwide): 3.8% at 30-34, 6.3% at 40-44, 10.2% at 50-54, 15.8% at 60-64, 23.4% at 70-74. The study starts at age 30; 0% at 18-24 and 1.0% at 25-29 are assumptions.
chronic kidney disease	18-44: 6.0%; 45-64: 13.0%; 65-100: 34.0%	3.4%	4.3%	6.1%	6.7%	14.5%	19.4%	19.3%	CDC. Chronic Kidney Disease in the United States, 2025 (NHANES August 2021-August 2023). 6% at 18-44, 13% at 45-64, 34% at 65+.

Condition	Published pattern	18-29	30-39	40-49	50-59	60-69	70-79	80-100	Source
osteoarthritis	18-29: 0.0%; 30-39: 2.0%; 40-49: 6.0%; 50-59: 14.0%; 60-69: 22.0%; 70-79: 30.0%; 80-100: 35.0%	0.0%	0.9%	2.9%	8.2%	11.2%	15.4%	19.7%	GBD 2021 Osteoarthritis Collaborators: 15% of people aged 30 and over, rising steeply with age. The age-group values are assumptions with that shape; 0% under 30.
atrial fibrillation	18-29: 0.0%; 30-54: 0.1%; 55-64: 1.3%; 65-74: 4.0%; 75-79: 7.3%; 80-100: 9.0%	0.0%	0.1%	0.1%	1.7%	3.9%	7.6%	14.0%	Go AS et al. Prevalence of diagnosed atrial fibrillation in adults (ATRIA study). JAMA 2001;285:2370-2375. About 0.1% under 55, rising to 9.0% at 80+. Middle age groups interpolated; 0% under 30 is an assumption.

Condition	Published pattern	18-29	30-39	40-49	50-59	60-69	70-79	80-100	Source
heart failure	18-29: 0.0%; 30-49: 0.2%; 50-69: 1.5%; 70-100: 10.0%	0.0%	0.2%	0.2%	0.4%	0.9%	6.4%	4.9%	McDonagh TA et al. 2021 ESC heart failure guidelines: 1-2% of adults, rising to 10% or more above age 70. Values under 70 are assumptions chosen to fit the 1-2% overall figure; 0% under 30.

Appendix A. Example Narrative

Narrative A-001 from the oncology trial, with every tagged item colored by its risk tier. The number after each tier name is the count of items of that tier in this narrative.

Narrative A-001, coloured by risk tier

Tier 1: direct (1) | Tier 2: high (3) | Tier 3: medium (2) | Tier 4: low (2) | Keep (5)

Subject 1003-0009 was a 66-year-old woman who took part in the study and whose occupation was member of parliament. The disease under study was non-small cell lung cancer, with a disease duration of 15 months. Her relevant medical history comprised migraine and osteoarthritis, and she was taking two concomitant medications, sumatriptan and paracetamol.

Study treatment was quelvatinib at a dose of 160 mg, administered by the oral route once daily, and the first dose was given on 04-Jul-2024. The participant developed grade 5 respiratory failure, which began on 14-Dec-2024 and was classified as a serious adverse event because it resulted in death. She was admitted to hospital because of worsening breathlessness. CT imaging showed growth of the lung tumour together with a large pleural effusion. A supporting laboratory result was an arterial oxygen tension of 6.9 kPa (normal range 10.5 to 13.5 kPa).

Her clinical course after admission to hospital was one of progressive decline. Despite drainage and non-invasive ventilation, her breathing continued to worsen, and she died in hospital on Study Day 176. No autopsy was done, and the stated cause of death was grade 5 respiratory failure from progression of the lung tumour. Action taken with the study drug: not applicable. Outcome: fatal.

The investigator's causality assessment was that the event was not related to the study drug. The investigator considered progression of the lung tumour to be the more likely cause of the event.

Appendix B. Rules for Tagging Items and Assigning Risk Tiers

Every person, site, investigator and drug in the narratives is fictional. This guide says how each piece of text in a narrative is marked, given a tier and given a weight. The aim is that anyone who follows it ends up with the same answer key. No second annotator tested this.

B.1 What counts as one item

An item is one piece of information that could point to a person, however many words it takes. "12 March 2025" is one item, not three. So are "Study Day 32", "91-year-old" and "Dr A. Kowalski". The tag covers the whole identifier, never part of it. If a longer phrase contains a smaller item, the longer phrase is the item. Tags never sit inside other tags.

B.2 Repeats

Every mention of the same item shares one item ID. Each mention is marked and counted as its own mention. All mentions of an item have the same tier and the same weight.

B.3 Tiers

Tiers come from how many people in the fictional population of 10,000 share the item. The first cut-off follows the 0.09 re-identification threshold: a value shared by 11 people has a count-based risk of 1/11 = 0.0909, which is just above 0.09, so 11 people or fewer is high risk. The cut-offs between tiers 3 and 4 (500 people) are our own choice. A script assigns every tier from a table of counts. IDs, staff names, calendar dates and study days have no count, so they use fixed rules. A column in the key records which rule set each tier. Table B1 lists the tiers, with an example from the data and the S2 and S3 weights for each.

Table B1. The risk tiers and the Keep class, with the rule for each, an example from the data, the number of people who share the example, and the S2 and S3 weights.

Tier	Name	Rule	Example from the data	People who share it	S2	S3
1	Direct	Points to one person on its own (fixed rule)	Subject 1008-0054	1	5	1
2	High	Shared by 11 people or fewer, or any calendar date (fixed rule)	FKRP-related limb-girdle muscular dystrophy	1	4	1
3	Medium	Shared by 12 to 500 people, or a study day (fixed rule)	Galway	74	2	0.0135
4	Low	Shared by more than 500 people	gastro-oesophageal reflux disease	1,285	1	0.000778
Keep	Utility	Clinical content that must stay	grade 5 respiratory failure	not counted	none	none

S2 gives fixed points per tier. S3 is 1 divided by the number of people who share the item, capped at 1. It is a rarity measure for a single value and not the risk of a whole narrative. Dates and study days take the median S3 weight of the counted items in their tier: 0.5 for dates and 0.006757 for study days.

B.4 Difficult cases

Five difficult cases are planted in the records: negation, second-hand claims, family history, abbreviations and mixed date styles. Each is marked like any other item, and the key carries a flag with the difficult case's name. A negated rare disease is still marked. "No history of" a rare disease still points to a family risk, so it keeps the tier of the disease and gets the flag negation. A date and a study day written next to each other are two items.

B.5 Keep items

Keep items are the clinical content that has to stay for the narrative to be useful: the adverse event, the study drug, key lab results and the trial's own disease, which is public from the trial registration. They have no tier and no weight. They are used for the keep rate and for over-removal. A tool that removes them makes the narrative less useful to a reader, even if it also removes every identifier.

B.6 How the key was checked

A script using only regular expressions and table lookups tagged and checked all 150 narratives. The author then checked 50 randomly chosen narratives (about a third, seed 20261020) by hand in Label Studio, with the draft key shown as pre-labels. Each tag and tier was confirmed, and identifiers written without a tag were to be added. None of the 50 narratives held an untagged identifier (leak rate 0%), and no correction was needed. The key was then frozen.

Appendix C. Standard Prompt for the Language Model

C.1 Prompt for the LLM tool

This is the standard prompt sent to qwen3:8b for every test narrative. The text "[plain narrative]" stands for the narrative with its tags removed. The model ran with temperature 0, seed 20261026, JSON output and a context of 4,096 tokens.

> Below is a patient narrative from a clinical trial report.
> List every piece of text that identifies the patient or could help identify them: names, ID numbers, dates, ages over 89, places and contact details.
> Copy each piece exactly as it appears in the text. List each piece once, even if it appears more than once.
> Return JSON only, in this form: {"remove": ["text 1", "text 2"]}
> Narrative:
> [plain narrative]