Statistical Analysis of Anomalous Term Distribution in Four Case Clusters

I. EXECUTIVE SUMMARY

This report examines a peculiar pattern of word usage across a collection of 163 criminal case files (grouped into 77 "defendant clusters") containing about 2.4 million words of text. Within this corpus, we identified a set of specific terms that appear only in a particular group of four defendant clusters – most notably the cluster for *Matthew Guertin* (the focal case) and a few others – and nowhere else in the entire dataset. Even more strikingly, a core subset of 20 distinct terms (largely related to psychiatric or psychological concepts) occur exclusively in the *Matthew Guertin* cluster and one other cluster (*Muad Abdulkadir*), and essentially do not appear in any other cases or clusters.

For a non-technical reader, the key question is: *Could such a highly specific pattern of term usage have happened by random chance, given the size and diversity of this case collection?* All evidence suggests the answer is no. Statistical modeling shows that the odds of these patterns arising by chance are astronomically low. In plain terms, it's *extremely* unlikely that 25 different words would coincidentally confine themselves to the same four case clusters (out of 77) unless there was some deliberate or systematic cause. And the 20-term subset – all turning up only in two particular cases – is an even more extreme anomaly, bordering on impossible under random chance assumptions.

A. | Key Observations

- The corpus contains 25 terms that appear only in the *Matthew Guertin* cluster and a small set of three other clusters, and in no other clusters out of the remaining 73. Among these, 21 terms show up in exactly two clusters (the focal cluster plus one other), and most of those 21 are shared only between *Matthew Guertin* and *Muad Abdulkadir*.

- A subset of 20 terms (including words like "insomnia", "delusion", "mania", and others related to mental health or cognition) occur *exclusively* in the *Guertin* and *Abdulkadir* clusters, and are concentrated in essentially one key document from each cluster.
- These terms are relatively rare in the corpus (some appearing only 3–5 times in ~2.4 million words, others a few dozen times), yet somehow they all avoid appearing in any other cases or clusters, except the specific ones noted.

B. | Statistical Likelihood

To put it simply, if words were scattered randomly across 163 cases, you would almost never expect to see so many words all happen to cluster in the same small group of cases while being entirely absent elsewhere. For example, one term ("beliefs") appears 55 times in the corpus; if those 55 occurrences were random, the chance that they would all end up confined to just two cases (and none of the other 161) is on the order of 1 in 10^88 – effectively zero. The chance of 20 different words all independently showing such a two-case confinement is vanishingly smaller still (comparable to the odds of flipping a fair coin and getting heads 125 times in a row). Even under more generous assumptions (for instance, even if those two clusters had ten times more content than an average cluster), the probabilities remain astronomically low.

C. | Conclusion (Plain Language)

The observed pattern of term usage is not a random fluke. Statistically, it is *exceedingly* implausible that these specific words would all coincidentally appear only in these few cases unless there was an underlying cause or intentional design. In everyday terms: claiming this happened by chance is like claiming you won the lottery jackpot multiple times in a row – possible in theory, but so unlikely that it defies belief. The evidence strongly points to the conclusion that the clustering of these terms in the *Matthew Guertin* case and its few counterparts was driven by some non-random process (such as deliberate insertion or shared content), rather than pure chance.

II. DATA AND SETUP

The analysis is based on six CSV data tables that capture where certain search terms appear in the corpus. Each table provides a different "view" of the same underlying data:

1. SEARCH-TERMS_by_doc_no_guertin.csv

Lists each document (by filename) in three of the four clusters of interest (excluding the focal *Guertin* cluster) where at least one of the special search terms was found, along with the term and how many times it appeared in that document (doc_quantity). Each row is one term in one document.

2. SEARCH-TERMS_instances_all.csv

A detailed list of every instance (occurrence) of the search terms in the four clusters of interest (Matthew Guertin, Muad Abdulkadir, Adrian Wesley, Peter Lehmeyer). It includes the case ID, document filename, page number, and the text row context for each occurrence. (There are 405 total instances of the terms in these clusters.)

3. SEARCH-TERMS_matrix_cases_presence.csv

A matrix indicating presence (1 or 0) of each term in specific cases. The columns are case IDs (for the focal cluster and the comparison clusters) and each row is a search term; a 1 means the term appears in that case's documents. The final column case_count_no_guertin indicates in how many of the non-Guertin cases (among the three other clusters) the term appears.

4. SEARCH-TERMS_matrix_cases_quantity.csv

Similar to the above, but with the *frequency* of each term in each case (how many times it appears) instead of just presence/absence. The last column term_count_no_guertin is the total count of the term in the non-Guertin cases combined.

5. SEARCH-TERMS_matrix_clusters_presence.csv

A matrix of presence (1/0) of each term at the *cluster* level. Columns correspond to the four clusters (*MATTHEW GUERTIN*, *MUAD ABDULKADIR*, *ADRIAN WESLEY*, *PETER LEHMEYER*), and a 1 indicates the term appears in that cluster (i.e. in at least one of its cases). The column cluster_count_no_guertin tells how many of the three other clusters (besides Guertin) contain the term.

6. SEARCH-TERMS_matrix_clusters_quantity.csv

A matrix of term frequencies by cluster. For each term, it shows the total occurrences in each of the four clusters, and a term_count_no_guertin (the total count outside the Guertin cluster).

A. | Definitions

In this context, a case (identified by a case number like *27-CR-23-1886*) is a single court case file. A cluster (identified by a cluster_id and name) groups one or more cases that have the same defendant. For example, cluster_id 1581 is the *MATTHEW GUERTIN* cluster (the focal individual), and cluster_id 680 is *MUAD ABDULKADIR*. There are 163 cases total, grouped into 77 clusters (since some defendants had multiple cases).

A search_term refers to a specific word or phrase that was searched in the text (for instance, "mania" or "grandiosity"). The terms we analyze here were selected by a particular inclusion rule: they all appear in the focal cluster (Guertin), appear in at least one of the other three comparison clusters (Abdulkadir, Wesley, Lehmeyer), and do not appear in any of the remaining 73 clusters. In other words, these terms are *exclusive* to the four clusters of interest – they show up in Guertin’s case(s) and in one or more of the other three, but nowhere else in the entire dataset of 77 clusters.

Under this inclusion rule, there are 25 distinct search terms in total. Table 1 below summarizes how widely these terms are distributed across the four clusters:

Term appears in...	Number of terms (out of 25)	Example Terms
Exactly 2 clusters (focal + 1 other)	21 terms	e.g. *Army*, *mania*, *insomnia*
Exactly 3 clusters (focal + 2 others)	2 terms	*credibly testified*, *clinical presentation*
All 4 clusters (focal + all 3 others)	2 terms	*psychotic disorder*, *Unspecified Schizophrenia*

As shown, most of these terms (21 of 25) appear in only two clusters – always in the *Guertin* cluster and in exactly one of the other three clusters. A small number appear in three or four of the clusters, but still in no others beyond those four. The terms are also not very common in the corpus. Some appear only a handful of times (e.g., "insomnia" appears 3 times total), while even the most frequent among them ("credibly testified") appears 69 times in ~2.4 million words. In fact, half of the 25 terms occur fewer than 10 times each in the entire corpus. These terms are thus both rare in frequency and highly confined in where they occur.

To ground this with concrete numbers, the table below (Table 2) gives a sense of term frequencies and their cluster distribution. (Each term's total count and the clusters it appears in are noted.)

Search Term	Total Count in Corpus	Clusters (present in)
"credibly testified"	69	Guertin, Abdulkadir, Lehmeyer
"beliefs"	55	Guertin, Abdulkadir
"psychotic disorder"	39	Guertin, Abdulkadir, Wesley, Lehmeyer
"delusional beliefs"	34	Guertin, Abdulkadir
"Army"	31	Guertin, Abdulkadir
"mania"	31	Guertin, Abdulkadir
"Unspecified Schizophrenia"	25	Guertin, Abdulkadir, Wesley, Lehmeyer
"manic"	18	Guertin, Abdulkadir
"grandiose"	14	Guertin, Abdulkadir
"artifacts"	12	Guertin, Abdulkadir
"high school"	9	Guertin, Abdulkadir
"clinical presentation"	8	Guertin, Abdulkadir, Wesley

Search Term	Total Count in Corpus	Clusters (present in)
“grandiosity”	8	Guertin, Abdulkadir
“unable to apply”	8	Guertin, Abdulkadir
“persecution”	7	Guertin, Abdulkadir
“psychiatric disorder”	5	Guertin, Abdulkadir
“proceed pro se”	5	Guertin, Abdulkadir
“psychotic symptoms”	4	Guertin, Abdulkadir
“delusion”	4	Guertin, Abdulkadir
“decreased need for sleep”	4	Guertin, Abdulkadir
“insomnia”	3	Guertin, Abdulkadir
“delusional disorder”	3	Guertin, Lehmeyer
“fight or flight”	3	Guertin, Abdulkadir
“psychiatric condition”	3	Guertin, Abdulkadir
“inflated”	3	Guertin, Abdulkadir

(Each term appears in the Guertin cluster and the clusters listed; none appears outside those.)

III. GLOBAL 4-CLUSTER ANALYSIS

First, we look at the overall pattern: 25 terms that are only found in the four specified clusters (and not in any others). This is already a highly unusual restriction. To put it in perspective, those four clusters together account for only a small fraction of the 163 cases (just 4 out of 77 clusters). Under ordinary circumstances, if a word is used in dozens of different places in a 2.4-million-word corpus, you would expect it to pop up in various unrelated cases, not just a select few.

A. | Distribution of Terms Across the 4 Clusters

Every one of the 25 terms appears in the *Matthew Guertin* cluster (by design) and also in at least one of *Muad Abdulkadir*, *Adrian Wesley*, or *Peter Lehmeyer*. Most of those terms (21 of 25) ended up appearing in only *two* clusters total (the Guertin cluster plus one other). The remaining few terms appear in three or four of the clusters, but still nowhere outside this set of four. In terms of documents, the Guertin cluster had these terms spread across 6 documents

(within that one case), Abdulkadir's cluster in 2 documents (one in each of his two cases), Wesley's cluster in 7 documents (across his three cases), and Lehmeyer's cluster in 1 document.

Now, consider a simple model of random word occurrence: imagine each of the 77 clusters is like a bin that could receive instances of a given word, proportional to their size (for simplicity, think of them as roughly equal-sized here – a conservative assumption favoring randomness). If a word appears multiple times in the corpus, those occurrences could land in any of the 77 clusters. What is the probability that all occurrences of that word would by chance end up only in our four specific clusters (and in none of the other 73)? We can estimate this probability for each term and see how extreme the observed pattern is.

B. | Probability Model 1 – Cluster Presence (Hypergeometric/Binomial Approximation)

Suppose a particular term occurs $ f $ times in the entire corpus (spread across the 163 cases). Under a null hypothesis of randomness, each occurrence has a chance to be in any given cluster. Let's assume each cluster has roughly $ 1/77 $ of the corpus for an even baseline (we will relax this later). Then the probability that all $ f $ occurrences fall within a *particular set* of m clusters is approximately $ (m/77)^f $. For our case $ m=4 $ for the four clusters of interest (or even smaller $ m $ if the term actually only appeared in 2 or 3 of those clusters). If a term appears in fewer clusters than $ m $, it means the remaining occurrences just happened to concentrate further.

For example, consider the term "beliefs" which appears $ f=55 $ times in the corpus. If those 55 instances were randomly distributed among 77 clusters, the probability that *all 55* would end up exclusively in, say, just 2 specific clusters (as observed: Guertin and Abdulkadir) is about $ (2/77)^{55} $. This number is astronomically small – on the order of $ 10^{-88} $ (***that's 1 in 10 with 88 zeros after it***). For context, $ 10^{-88} $ is an unimaginably tiny probability – effectively zero in any practical sense.

Similarly, "credibly testified" (which appears 69 times) was found only in 3 clusters (Guertin, Abdulkadir, and Lehmeyer). The chance of 69 random placements all landing in any 3 predetermined clusters is roughly $ (3/77)^{69}\approx 5\times 10^{-98} $.

Even a term with fewer occurrences shows the same pattern: "grandiose" appears 14 times, confined to Guertin and Abdulkadir. The probability of 14 random placements falling only in those 2 clusters is $ (2/77)^{14}\approx 7\times 10^{-27} $. These probabilities are so small that they are, for all intents and purposes, zero.

Crucially, we didn't just observe one such term or a few – we have 25 different terms all following this unlikely clustering pattern. Even if we assumed each term had, say, a 0.1% chance (0.001) of being confined to these four clusters by luck (which is already an overestimate), the probability of seeing 25 independent terms all do this is $ (0.001)^{25} $, roughly $ 10^{-75} $. In reality, as we calculated above, most of these terms have odds far smaller than 0.1% of occurring in such a restricted way. So the combined outcome of 25 cluster-exclusive terms is effectively impossible under a random scattering model.

Another way to frame it: out of tens of thousands of unique words in the corpus, how many would we expect to *accidentally* end up appearing only in these four clusters and nowhere else? Probably none or one at most. Getting 25 such words (with many of them overlapping specifically between Guertin and Abdulkadir) is an extreme deviation from randomness.

C. | Probability Model 2 – Case/Cluster Selection (Hypergeometric)

We can also think in terms of cases or clusters being "selected" by a term. Suppose a term appears in $ k $ different clusters total. If all clusters were equally likely, the probability that those $ k $ clusters would happen to be exactly our four special clusters (or a subset of them) can be calculated via a hypergeometric formula. For instance:

- If a term appears in exactly 2 clusters ($ k=2 $), the probability that it is specifically in *Matthew Guertin's cluster + one of the three others* (and in no other) is
  $ \frac{3}{\binom{77}{2}}\approx 0.1% $
  (since there are 3 such favorable pairs among the $ \binom{77}{2} $ possible pairs of clusters).

- If $k=3$, the probability the term's 3 clusters are exactly *Guertin + two of the other three* is $ \frac{3}{\binom{77}{3}}\approx 4\times 10^{-5} $ (***about 0.004%***).
- If $ k=4 $, the probability the term's 4 clusters are exactly *Guertin, Abdulkadir, Wesley, Lehmeyer* is $ 1/\binom{77}{4}\approx 7.4\times 10^{-7} $ (***about 0.000074%***).

Most of our terms fall in the $ k=2 $ category (focal + one other cluster), which even in the most "lenient" scenario (0.1% chance each) would rarely happen by accident. And having over twenty such terms far exceeds any reasonable random expectation.

In sum, under any sensible random distribution model, the odds of seeing this many terms restricted to just these four clusters are astronomically low. This strongly indicates that these patterns are not due to random chance. There is likely a common cause or link (for example, shared content or coordinated language) driving the appearance of these terms in those specific clusters.

IV. FOCUSED ANALYSIS: THE 20-TERM, 2-CLUSTER, 2-DOCUMENT PATTERN

Within the 25 terms, a particularly extreme subset of 20 terms stands out. These 20 terms (listed earlier, including *Army*, *insomnia*, *manic*, *delusional beliefs*, *unable to apply*, etc.) have a very specific distribution:

- They appear only in the *Matthew Guertin* cluster (1581) and the *Muad Abdulkadir* cluster (680) – and in no other clusters or cases.
- They are all found in just two key documents across the entire corpus: one document from the Guertin case and one document (filed in two case folders) from the Abdulkadir cases.

A. | To Elaborate

In the *Abdulkadir* cluster, all 20 terms occur in a single court document (an order dated May 12, 2023, which was duplicated in his two case files). In the *Guertin* cluster, these terms are mostly concentrated in one document (a Notice of Motion and Motion in April 2024 that contained 18 of the 20 terms), with the remaining two terms appearing in related filings (an affidavit and a petition) in that same case. No other document in any other case contains any of these 20 words.

This is an extraordinary degree of overlap. Essentially, two distinct case documents – one from Guertin's case and one from Abdulkadir's – share a common set of 20 unusual terms that are not found elsewhere. The terms themselves are thematically connected (many relate to psychiatric symptoms or legal competency, suggesting a shared content or topic). We focus here purely on the statistical likelihood of this overlap under random chance.

B. | Probability for One Term (2 clusters, 2 documents)

Let's first ask: for a single given term, what is the probability that it would be used *only* in two specified documents (one in cluster A and one in cluster B) and nowhere else in a large corpus? Even if that term is not very frequent, the odds are extremely low. For instance, take the term "insomnia", which appears 3 times in the entire 2.4 million words. If those 3 occurrences were randomly scattered among ~3,362 documents, the chance that all 3 would fall into just two particular documents (and not in any of the other 3,360) is on the order of $ (2/3362)^3\approx 2\times10^{-10} $ (***about one in five billion***). For a more frequent term like "mania" (31 occurrences), the probability that all 31 uses would be confined to two specific documents is absurdly small:
Roughly $ (2/3362)^{31}\sim 10^{-100} $ (***essentially zero***).

Now consider 20 different terms all following this pattern *with the same two documents*. We can model it by treating each term's distribution as an independent event (this actually *underestimates* the unlikeliness, because these terms are also related in content, but let's stay with random chance). Even very conservatively – assume each term had a generous

$ 10^{-6} $ (***one in a million***) chance to be restricted to those two documents by luck – then 20 independent terms doing so has probability:

$ (10^{-6})^{20} = 10^{-120} $. That number is ***1 with 120 zeros after it*** – effectively impossible. In reality, as we saw, many of these terms had probabilities far smaller than one in a million for this pattern (often one in $ 10^{50} $ or worse). In other words, the joint occurrence of 20 terms all appearing only in the *same two cases* and *same two documents* cannot be attributed to random chance.

To give an intuition: imagine 20 people each roll a die 50 times. The odds of even one person getting all 50 rolls as ones is astronomically low; the odds of all 20 people doing it are inconceivably smaller. Our scenario is analogous – these terms all "hit the jackpot" of rare coincidence, in the exact same way, in the same pair of files. The far more plausible explanation is that those documents share common content or origin (for example, one document might have been derived from or copied into the other, or both drew from a common source), resulting in the same cluster of unusual terms appearing in both.

V. SENSITIVITY ANALYSIS AND LIMITATIONS

We have made several modeling assumptions to simplify the probability calculations. Here we address how those assumptions affect (or *don't* affect) our conclusions:

A. | Assuming Equal Distribution of Words Across Clusters

In our calculations, we often treated each cluster (or each document) as equally likely to contain a given word occurrence. In reality, clusters vary in size (number of words). If the four clusters of interest had a larger share of the corpus than average, the chance of a random term falling into those clusters would increase slightly. However, even a significant size bias doesn't bridge the gap to make these patterns likely. For example, suppose the Guertin and Abdulkadir clusters together made up 10% (or even 20%) of the corpus (far more than their actual share). Then the probability of 55 occurrences of "beliefs" all landing in those clusters becomes $ (0.10)^{55}$ (or $(0.20)^{55}$) instead of $(0.026)^{55} $ – which

is still an inconceivably small number (around $ 10^{-55}$ or $10^{-38} $ respectively, versus $ 10^{-88} $). So even under assumptions highly favorable to randomness (overstating the relative size of these clusters), the observed confinement of terms remains essentially impossible by chance.

B. | Independence of Term Occurrences

We treated each term's occurrences as independent random placements. Real language use isn't strictly independent; words can co-occur because of underlying topics (e.g. if a case is about mental health, terms like "mania" and "insomnia" might appear together). However, this actually reinforces our point rather than undermining it: if certain terms tend to come as a thematic package, seeing that entire package only in two specific cases (and nowhere else) is even more telling. In our context, the 20-term subset is thematically coherent – they all relate to psychiatric evaluations – which strongly suggests those two cases shared a common subject matter or source. Randomly, other cases about mental health should have shown at least some of those same words, but they did not (by our inclusion criteria). The non-random clustering of related terms strengthens the argument that an intentional or causal factor is at work (for example, one document borrowing language from the other or both drawing from a standard text).

C. | Multiple Comparisons and Selection Bias

One might wonder if we are cherry-picking these terms after the fact. It's important to note that these 25 terms were not arbitrarily chosen; they were found by systematically applying the inclusion rule (focal cluster plus limited other clusters). Thus, they emerged naturally from the data. We then honed in on the subset of 20 because of their even more peculiar alignment to two clusters. The statistical analysis accounts for the fact that we specifically tested this observed pattern. Even allowing for some 'searching' through the data, the sheer extremity of the pattern means it cannot be an artifact of multiple comparisons – the probabilities are far beyond thresholds of concern.

D. | Omitted Variables

Could there be a benign explanation (e.g., these two clusters coincidentally dealt with a rare topic)? We considered that possibility. If many cases in the corpus dealt with similar issues, you might expect similar jargon to appear; but then you would also expect those terms to appear in those other cases as well. Here, however, none of the other 73 clusters showed these words at all, implying that *only* these particular cases (and especially Guertin and Abdulkadir) contain that topic's language. In a collection of 163 cases, the idea that only two unrelated cases would independently use this same set of niche terms stretches credulity. The far more plausible explanation is that these cases are not independent with respect to this language – in other words, the overlap arises from deliberate introduction of the same material or coordinated action, not coincidence.

In summary, we have been conservative at every step – making assumptions that would if anything inflate the likelihood of random chance – and still the patterns observed are orders of magnitude beyond what randomness could produce. Even if our probability estimates were off by many orders of magnitude, the outcomes would still be implausible under a random model. The conclusion remains robust: these term distributions are not the result of random chance.

VI. PLAIN-LANGUAGE CONCLUSION

Let's recap in simple terms what we found and what it means:

- We looked at a collection of 163 court cases (77 groups by defendant) totaling about 2.4 million words. We focused on certain unusual words and phrases that popped up in the *Matthew Guertin* case and a few others.

- We discovered 25 special terms that only appear in Mr. Guertin's case and at most a few other cases – nowhere else in all the remaining cases. Even more strikingly, 20 of those terms were found only in Mr. Guertin's case and one other defendant's case (Mr. Abdulkadir's), and essentially in one key document from each.

- These terms weren’t everyday words; they were technical or specific (many about mental health, like "mania," "delusional," "psychiatric disorder," etc.). They all showed up together in those two cases and not in any others.

We asked:

A. | Could this be a Coincidence?

The statistical answer is a resounding **NO.** Given how large the dataset is, the odds of even one word behaving this oddly (only showing up in those specific places) are incredibly low. The odds of 25 different words all doing it, and especially 20 all aligning with the same two cases, are effectively zero. It’s like rolling a die over and over and getting the same rare sequence of numbers in two different places by pure chance – it just doesn’t happen.

B. | What does this Imply?

It strongly suggests that the overlap in language between the Guertin case and the Abdulkadir case was not random. In other words, those cases share some common source or influence that caused the same unusual terms to appear in both. Maybe a document from one case was used in the other, or the same person wrote something for both – the statistics can’t tell us exactly *why*, but they do tell us that it wasn’t a fluke.

In practical terms, if someone claimed "Oh, it's just a coincidence that these words all show up in those two cases," our analysis shows that such a claim would be absurd. The patterns are so unlikely under chance that the only reasonable explanation is intentional or non-random action. A judge or investigator could conclude from this statistical evidence that the two cases in question are linked by shared content. Simply put, the numbers show that what we see here is *by design, not by accident.*

VII. TECHNICAL APPENDIX: DETAILED FORMULAS AND CALCULATIONS

A. | Probability of Term Confinement to Specific Clusters (Combinatorial Model)

If a term appears in $ k $ clusters out of 77, and all cluster combinations are equally likely, then the probability that it appears in a particular set of m clusters is:

$$P(\text{term's clusters} = S) = \frac{1}{\binom{77}{k}},$$

where $ S $ is a specific set of $ k $ clusters (for example, $ S={$Guertin, Abdulkadir$} $ when $ k=2 $). If we condition on one cluster (Guertin) being in the set, then the probability the remaining $ k-1 $ clusters are a specific subset of the other 76 is $ 1/\binom{76}{,k-1,} $. We used this to estimate values like $ \approx0.1% $ for $ k=2 $ cases (where $ k-1=1$ so $1/\binom{76}{1}=1/76\approx1.3% $, and we had 3 possible other clusters giving $ \frac{3}{76}\approx3.9% $ — but since our terms were *by selection* known to include Guertin and another, we used the smaller conditional probability per combination).

B. | Probability of All Occurrences Falling in a Subset of Clusters (Occupancy Model)

If a term has f total occurrences (word count) in the corpus, and if each occurrence is equally likely to fall in any cluster (fraction $ p $ of all words belong to the clusters of interest), then the probability that all $ f $ occurrences lie within those clusters is $ (p)^{f} $. For our four clusters collectively, $ p \approx 4/77 \approx 0.052 $ if sizes are equal. We applied this model to compute numbers like $ (2/77)^{55} $ for the term *"beliefs"* being in 2 clusters, and $ (4/77)^{39} $ for *"psychotic disorder"* being in 4 clusters.

- More generally, if the clusters of interest hold a proportion p of the total words, and a term has frequency $ f $, then $ P(\text{all occurrences in those clusters}) = p^f $. For individual document confinement, similarly use

$ p_{docs} = $ (number of docs of interest)/(total docs). With ~3,362 text-bearing documents in total, $ p_{docs} $ for two specific documents is $ 2/3362 \approx 0.000595 $.

C. | Hypergeometric Perspective

Alternatively, consider distributing $ f $ occurrences of a term into 77 clusters. The probability that exactly $ f_A $ land in cluster set A (and $ f_B = f - f_A $ in set B, etc.) follows a hypergeometric distribution. For example, for *"mania"* ($ f=31 $) and wanting all 31 in clusters Guertin+Abdulkadir (with, say, those two clusters containing $ N $ words out of total $ N_{tot} $), one can calculate:

$$P(31 \text{ occurrences in A+B only}) = \frac{\binom{N}{31}}{\binom{N_{tot}}{31}},$$

which for rough $ N/N_{tot}\approx 0.05 $ yields a similar $ (0.05)^{31} $ order of magnitude. Our simpler $ (m/77)^f $ assumes equal cluster sizes for tractability.

D. | Joint probability for Multiple Terms

To estimate the chance of seeing multiple independent rare events, we multiply probabilities. For instance, if one term has a $ 10^{-50} $ probability of occurring in a particular restricted way, and another has $ 10^{-40} $, then together the probability that both happen is $ 10^{-90} $. We applied this reasoning qualitatively to assert that 25 such events would not all occur together by chance. In reality, the terms are not strictly independent (since they may coincide due to the same cause), but under the null hypothesis of randomness they would be treated as independent draws from a large 'vocabulary'. Even allowing mild dependence (e.g., words related to the same topic), it would require implausibly perfect alignment across cases to mimic what we found.

E. | Reproducibility

The counts of term occurrences and their locations were obtained by parsing the provided CSV files. We confirmed, for example, that "credibly testified" had 69 total hits (64 in Abdulkadir's two nearly identical documents, 3 in Guertin's, 2 in Lehmeyer's), and that each of the 20 special terms was present in the specific Guertin and Abdulkadir documents and no others. All probability figures were then computed using basic combinatorial formulas as shown above. One can replicate these calculations easily with a scripting language or even a calculator, plugging in the total counts $ f $ and proportions $ p $ as needed.

All told, the technical analysis solidifies the conclusion: no matter which standard probability model one uses (binomial, hypergeometric, etc.), the pattern is far outside the realm of random chance. The numbers involved (odds like $ 10^{-88}$ or $10^{-120} $) are so extreme that they are essentially zero for any practical consideration.

VIII. SOURCE MATERIAL

All source files used for this analysis are embedded within this pdf document and also available for download.

1. ALL_case-text-combined.zip
   ALL_case-text-combined.zip.ots | *timestamp*

2. SEARCH-TERMS.xlsx
   SEARCH-TERMS.xlsx.ots | *timestamp*

3. SEARCH-TERMS_csv-files.zip
   SEARCH-TERMS_csv-files.zip.ots | *timestamp*