

When Documents Disagree: Measuring Institutional Variation in Transplant Guidance with Retrieval-Augmented Language Models

Yubo Li, MS¹, Ramayya Krishnan, PhD¹, Rema Padman, PhD¹
¹ Carnegie Mellon University, Pittsburgh, PA, USA

Abstract

Patient education materials for solid-organ transplantation vary substantially across U.S. centers, yet no systematic method exists to quantify this heterogeneity at scale. We introduce a framework that grounds the same patient questions in different centers' handbooks using retrieval-augmented language models and compares the resulting answers using a five-label consistency taxonomy. Applied to 102 handbooks from 23 centers and 1,115 benchmark questions, the framework quantifies heterogeneity across four dimensions: question, topic, organ, and center. We find that 20.8% of non-absent pairwise comparisons exhibit clinically meaningful divergence, concentrated in condition monitoring and lifestyle topics. Coverage gaps are even more prominent: 96.2% of question-handbook pairs miss relevant content, with reproductive health at 95.1% absence. Center-level divergence profiles are stable and interpretable, where heterogeneity reflects systematic institutional differences, likely due to patient diversity. These findings expose an information gap in transplant patient education materials, with document-grounded medical question answering highlighting opportunities for content improvement.

1 Introduction

Large language models (LLMs) are increasingly used to answer medical questions and support patient education, often through retrieval-augmented generation (RAG), in which responses are grounded in external documents such as clinical guidelines or institutional patient education materials. While prior work has focused on improving retrieval quality and reducing hallucinations, less attention has been paid to an upstream issue: can the underlying source documents themselves contain differing guidance? If so, what are the similarities and differences between handbooks on specific topics and what is their provenance?

This question is especially relevant in solid-organ transplantation, where substantial center-level variation is well documented. Analyses of national registry data show that the probability of receiving a deceased-donor kidney transplant within three years of waitlisting varies dramatically across centers within the same donation service area [1]. These institutional differences extend to patient-facing information: transplant center websites provide incomplete and inconsistent recipient selection criteria [2], patient education materials vary widely in readability and quality [3, 4], and a comparative analysis of transplant handbooks using NLP and generative methods found significant variation in the availability and interpretation of clinical guidance across centers [5].

These findings raise a critical question: when patients ask the same or a similar question to different centers' patient handbook documents, does the resulting advice differ in clinically meaningful ways? If so, content selection for patient handbooks has implications for both clinical and management decision making. In this work, we address this problem by grounding the same patient questions in handbooks from different U.S. transplant centers and systematically comparing the resulting responses. Our goal is not to identify a single correct answer but to measure the extent and structure of institutional variation across organs, topics, and centers, contributing to broader efforts to understand the information gaps in transplant patient education within the U.S. transplantation system.

2 Methods

2.1 Data Sources and Processing

Our study draws on two primary data sources: a corpus of transplant patient education handbooks that were generously shared by U.S. transplant centers and assembled by the non-profit Transplants.org for analysis, and a curated benchmark of patient questions extracted from multiple patient forums, spanning five solid-organ types. This section describes the collection, scope, and characteristics of each.

Transplant Patient Handbooks. We obtained a corpus of 102 patient education handbooks from 23 major U.S. solid-organ transplant centers, representing 16 of the nation's 20 largest programs by volume. The corpus spans five organ

types — heart (26), lung (26), kidney (22), liver (17), and pancreas (11) — and the contributing centers are geographically distributed across the United States (Figure 1), covering both large academic medical centers and community-based transplant programs. All documents were obtained as PDFs from the institutions. Table 1 summarizes the corpus by organ type.

Centers vary in how they organize patient education materials: some provide separate documents for the pre-transplant phase (evaluation, listing, and waiting) and the post-transplant phase (recovery, medications, and long-term follow-up), while others issue a single combined handbook. We treat each phase-specific document as a distinct unit, yielding 37 pre-transplant, 39 post-transplant, and 26 combined handbooks. Each is assigned a unique identifier encoding organ type, institution, and care phase.



Figure 1: Center distribution map for the handbook corpus. The 23 contributing U.S. transplant centers are geographically dispersed across the country.

Transplant Patient Question Set. We curated a benchmark of 1,115 patient questions to serve as the evaluation set for cross-center comparison. Questions were collected from diverse sources reflecting the real information needs of transplant patients and caregivers. Source types include healthcare institution Q&A pages (31.2%), community forums such as Reddit and Mayo Clinic Connect (25.1%), medical organizations including the National Kidney Foundation and the American Liver Foundation (24.9%), and other sources including government health agencies and patient advocacy sites (18.8%).

Table 1: Summary of the transplant handbook corpus by organ type. *Centers* indicates the number of distinct institutions contributing handbooks for each organ.

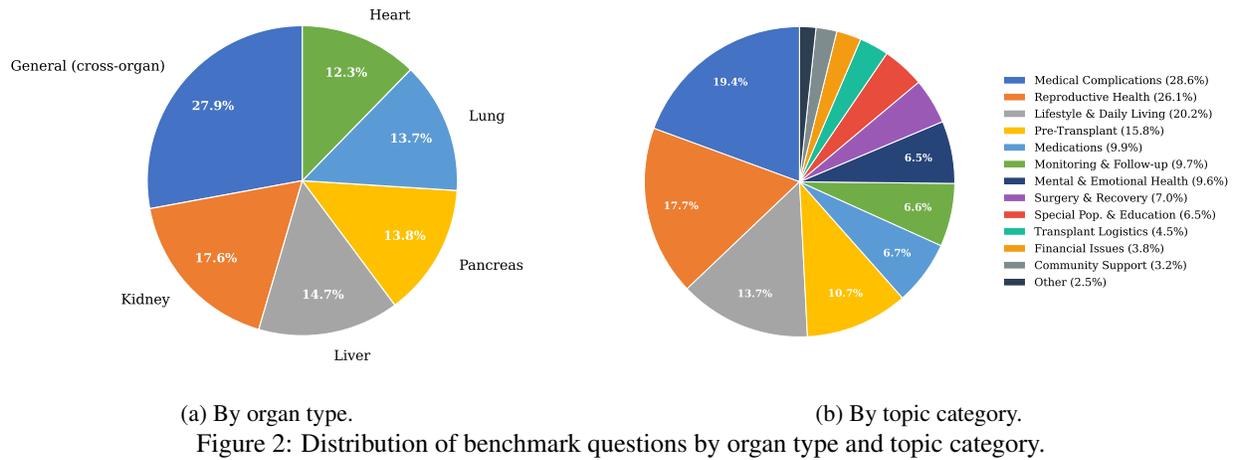
	Heart	Kidney	Liver	Lung	Pancreas
Handbooks	26	22	17	26	11
Centers	17	14	11	15	8
Pre-transplant	10	8	5	10	4
Post-transplant	11	10	4	11	3
Combined	5	4	8	5	4

Each question is annotated with: (i) an organ type label (heart, kidney, liver, lung, pancreas, or general), (ii) one or more clinical topic categories drawn from a 13-topic taxonomy, and (iii) fine-grained sub-topic tags (43 unique sub-topics). As shown in Figure 2b, the 13 topic categories are: *Medical Complications* (28.6% of topic annotations), *Reproductive Health* (26.1%), *Lifestyle & Daily Living* (20.2%), *Pre-Transplant* (15.8%), *Medications* (9.9%), *Monitoring & Follow-up* (9.7%), *Mental & Emotional Health* (9.6%), *Surgery & Recovery* (7.0%), *Special Populations & Education* (6.5%), and four smaller categories covering transplant logistics, financial issues, and community support. Questions are multi-labeled: a single question may be annotated with more than one topic and sub-topic to reflect cross-cutting concerns.

General-type questions account for the largest share of the benchmark (27.9%; Figure 2a) and address topics that span organ types, such as immunosuppressant side effects, reproductive health after transplantation, and mental health. At generation time, these questions are answered by *every* handbook in the corpus, while organ-specific questions are answered only by handbooks of the matching organ type. This design ensures comprehensive cross-center coverage for broadly relevant topics while maintaining clinical specificity for organ-level questions.

Questions were lightly paraphrased for anonymization and to ensure they are self-contained (i.e., interpretable without conversational context). The original sources were predominantly U.S.-based (69.9% geolocated to the United States), consistent with the U.S. institutional focus of the handbook corpus. The benchmark was drawn from a larger pool of over 3,000 candidate questions collected across organ-specific sources, which were filtered for relevance, deduplica-

tion, and quality to produce the final set of 1,115.



2.2 Document Extraction

Figure 3 illustrates the end-to-end pipeline. Raw PDF handbooks are converted to structured JSON representations using LlamaParse [6], a document parsing service that preserves section headings, paragraph boundaries, and page metadata. The output for each handbook is a JSON object containing: organ type, center name, care phase, source file path, full text, and a list of sections—each with its heading, body text, and page numbers. This structured extraction enables section-aware chunking in the downstream retrieval stage (Section 2.3). The extraction pipeline is idempotent: already-processed files are detected and skipped, ensuring resume-safe execution across incremental corpus updates.

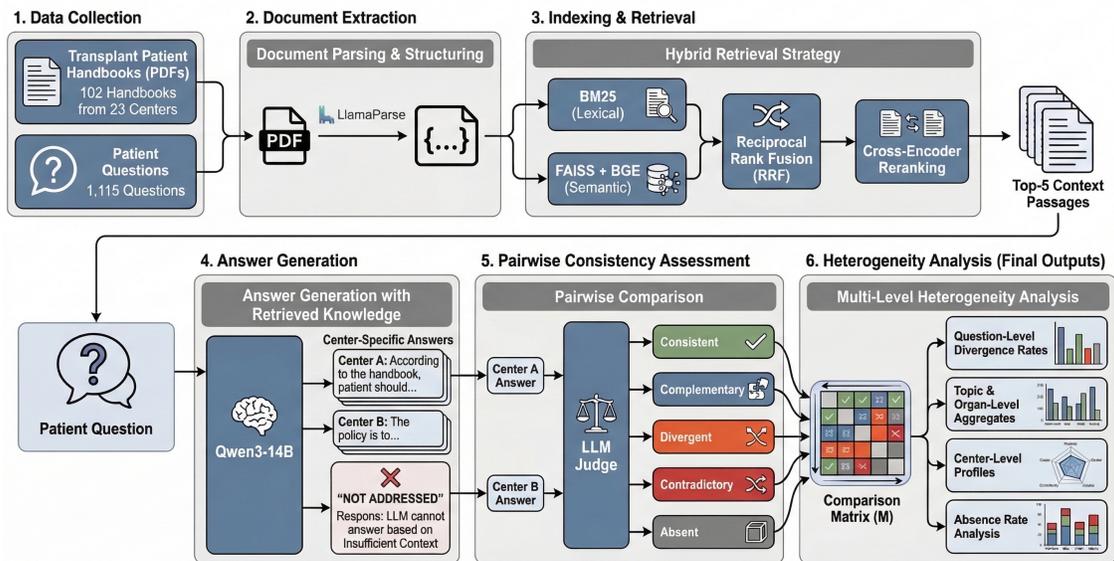


Figure 3: Overview of the experimental pipeline.

2.3 Retrieval and Answer Generation

To ground LLM-generated answers in center-specific content, we adopt a hybrid retrieval strategy that combines sparse lexical matching with dense semantic retrieval, followed by neural reranking.

Indexing. Leveraging the structured JSON output from the extraction stage (Section 2.2), each handbook is segmented into chunks at section boundaries. Sections exceeding 512 tokens are further split at sentence boundaries, with each sub-chunk inheriting the parent section heading as a prefix to preserve topical coherence. All chunks are indexed in two parallel structures: a BM25 [7] inverted index for sparse lexical retrieval and a FAISS [8] vector index using the BAAI/bge-large-en-v1.5 dense encoder [9] for semantic retrieval.

Hybrid Retrieval with Reranking. At query time, both retrievers independently return ranked candidate lists. These are merged using Reciprocal Rank Fusion (RRF) [10] with $k_{\text{RRF}} = 60$, which balances the complementary strengths of lexical and semantic matching. The fused candidate set is then reranked with a cross-encoder [11]), and the top-5 passages are selected as the retrieval context. This hybrid approach has been shown to outperform either retrieval modality alone on biomedical text [12].

Answer Generation. For each question–handbook pair, the top-5 retrieved passages are provided alongside the original question to the generation model (Qwen3-14B at temperature 0), which produces a grounded answer. The model is instructed to rely exclusively on the provided context and to return a standardized NOT ADDRESSED response if the handbook does not contain information relevant to the question, rather than hallucinate an answer.

2.4 Heterogeneity Analysis

Given the set of center-specific answers generated for each question, we analyze cross-center heterogeneity at multiple levels of granularity. We first identify coverage gaps through absence detection, then classify the relationship between every pair of center responses, and finally aggregate divergence and consistency metrics at the question, topic, organ, and center levels.

2.4.1 Absence Detection and Coverage Analysis

Before comparing center responses, each answer is screened for *absence*—whether the handbook failed to address the question. A fast heuristic first checks for the canonical NOT ADDRESSED prefix produced by the generation model. Answers not matching this prefix are passed to a secondary LLM-based binary classifier (Qwen3-14B) that determines via a structured YES/NO prompt whether the response substantively indicates non-coverage. Absence status is cached per handbook–question pair to avoid redundant inference.

For each organ type and topic category, we report the *absence rate*: the fraction of question–center pairs for which the handbook did not address the question. Let \mathcal{Q}_g denote the set of questions in group g (an organ type or topic category), and let $\mathcal{C}(q)$ denote the set of centers whose handbooks are queried for question q . The absence rate for group g is:

$$r_{\text{abs}}(g) = \frac{\sum_{q \in \mathcal{Q}_g} \sum_{c \in \mathcal{C}(q)} \mathbf{1}[\text{absent}(q, c)]}{\sum_{q \in \mathcal{Q}_g} |\mathcal{C}(q)|} \quad (1)$$

where $\mathbf{1}[\text{absent}(q, c)]$ is an indicator for whether center c ’s handbook does not address question q . High absence rates indicate systematic coverage gaps—topics that patients ask about but that institutional materials do not address.

2.4.2 Pairwise Consistency Assessment

Five-Label Comparison Taxonomy. For each pair of non-absent answers from two different centers, an LLM judge classifies their relationship into one of five categories. Table 2 defines each label and provides an illustrative example.

Pairs involving at least one absent answer are assigned ABSENT without further LLM inference. For all remaining (non-absent) pairs, the LLM judge is prompted with the original question and both center answers, and returns a structured JSON object containing: the classification label, a 2–3 sentence clinical justification, a divergence sub-topic tag (if applicable), and a clinical significance rating (low, medium, or high) for pairs labeled DIVERGENT or CONTRADICTORY. Inference is performed with greedy decoding (temperature 0) for reproducibility.

Comparison Matrix. For a question answered by N centers, the $\binom{N}{2}$ unique pairwise comparisons form a symmetric $N \times N$ comparison matrix \mathbf{M} , where entry M_{ij} records the label assigned to the pair of handbooks i and j . Diagonal entries are defined as CONSISTENT by convention. Each pairwise result is persisted independently as a JSON file, enabling resume-safe incremental execution over the full corpus.

Table 2: Five-label taxonomy for pairwise comparison of center-specific answers.

Label	Definition	Example
ABSENT	One or both answers indicate the handbook does not address the topic.	Center A provides dietary guidance; Center B’s handbook contains no relevant section.
CONSISTENT	Both answers provide the same clinical recommendation.	Both centers advise avoiding grapefruit due to tacrolimus interactions.
COMPLEMENTARY	Clinically compatible but differing in detail or scope.	Center A lists side effects; Center B additionally describes management strategies.
DIVERGENT	Substantive, clinically meaningful differences (e.g., different thresholds or timelines).	Center A recommends exercise at 6 weeks post-transplant; Center B at 8–12 weeks.
CONTRADICTORY	Directly opposing clinical guidance.	Center A allows ABO-incompatible live donors; Center B states they cannot proceed.

2.4.3 Question-Level Heterogeneity

For each question q , we quantify cross-center agreement and disagreement using two complementary metrics computed over all $\binom{N}{2}$ center pairs.

The *divergence rate* measures the fraction of non-absent pairs exhibiting clinically meaningful disagreement:

$$r_{\text{div}}(q) = \frac{|\{(i, j) : M_{ij}^{(q)} \in \{\text{DIVERGENT}, \text{CONTRADICTORY}\}\}|}{|\{(i, j) : M_{ij}^{(q)} \neq \text{ABSENT}\}|} \quad (2)$$

The *consistency rate* measures the fraction of non-absent pairs in full agreement:

$$r_{\text{con}}(q) = \frac{|\{(i, j) : M_{ij}^{(q)} = \text{CONSISTENT}\}|}{|\{(i, j) : M_{ij}^{(q)} \neq \text{ABSENT}\}|} \quad (3)$$

where $M_{ij}^{(q)}$ is the comparison label for question q between centers i and j . Questions with high r_{div} identify topics where institutional guidance is most fragmented, while questions with high r_{con} indicate consensus across responding centers. Note that $r_{\text{div}}(q) + r_{\text{con}}(q) \leq 1$, with the residual fraction accounted for by COMPLEMENTARY pairs.

2.4.4 Topic- and Organ-Level Aggregation

Both metrics are aggregated by topic category and organ type. For a group g (topic or organ) with associated question set \mathcal{Q}_g , the group-level divergence and consistency rates are:

$$R_{\text{div}}(g) = \frac{1}{|\mathcal{Q}_g|} \sum_{q \in \mathcal{Q}_g} r_{\text{div}}(q), \quad R_{\text{con}}(g) = \frac{1}{|\mathcal{Q}_g|} \sum_{q \in \mathcal{Q}_g} r_{\text{con}}(q) \quad (4)$$

We additionally report the proportion of questions within each group for which $r_{\text{div}} > 0$ (i.e., at least one divergent or contradictory pair exists). This two-metric approach distinguishes between groups where divergence is pervasive (many questions affected) and groups where it is concentrated in a few high-disagreement questions.

2.4.5 Center-Level Profiles

For each center c , we compute a heterogeneity profile by aggregating pairwise labels across all questions and partner centers. Let $\mathcal{P}(c)$ denote the set of all non-absent pairwise comparisons involving center c . The center-level divergence and consistency rates are:

$$R_{\text{div}}(c) = \frac{|\{(q, j) \in \mathcal{P}(c) : M_{cj}^{(q)} \in \{\text{DIVERGENT}, \text{CONTRADICTORY}\}\}|}{|\mathcal{P}(c)|} \quad (5)$$

$$R_{\text{con}}(c) = \frac{|\{(q, j) \in \mathcal{P}(c) : M_{cj}^{(q)} = \text{CONSISTENT}\}|}{|\mathcal{P}(c)|} \quad (6)$$

Centers with consistently high R_{div} across topics may reflect systematically different institutional policies, while centers with high R_{con} relative to peers indicate alignment with prevailing practice norms. These profiles enable identification of outlier institutions whose guidance departs most from the cross-center consensus.

3 Results

We applied the full pipeline to the corpus of 102 handbooks from 23 centers across five organ types, using all 1,115 benchmark questions. This section reports heterogeneity findings organized by global label distribution, coverage gaps, organ- and topic-level divergence, center-level profiles, and illustrative matrix visualizations.

3.1 Global Label Distribution

Across 1,115 questions and 102 handbooks, the pairwise comparison pipeline produced 1,772,261 handbook pairs. Of these, 1,704,242 (96.2%) were classified as ABSENT, reflecting the expected sparsity: most handbooks address only a subset of the benchmark questions. Among the 68,019 non-absent pairs (i.e., pairs where both handbooks substantively addressed the question), the label distribution was: COMPLEMENTARY 44,870 (66.0%), DIVERGENT 14,132 (20.8%), CONSISTENT 8,874 (13.0%), and CONTRADICTORY 143 (0.2%). Table 3 summarizes the global distribution.

Table 3: Global pairwise label distribution across all 1,115 questions. Percentages in the right column are computed over non-absent pairs only.

Label	Count	% of non-absent
ABSENT	1,704,242	—
COMPLEMENTARY	44,870	66.0%
DIVERGENT	14,132	20.8%
CONSISTENT	8,874	13.0%
CONTRADICTORY	143	0.2%
Total non-absent	68,019	100%

The dominance of COMPLEMENTARY labels indicates that when two centers both address a question, they most frequently provide compatible but differently scoped information. However, one in five non-absent pairs exhibits clinically meaningful divergence, and a small but non-negligible number involve direct contradictions.

3.2 Coverage Gaps: Absence Rates by Organ and Topic

Absence rates varied substantially across organ types and topic categories (Table 4). General-type questions, which are posed to all 102 handbooks, had the highest absence rate (90.5%), reflecting the fact that organ-specific handbooks rarely cover cross-cutting topics such as reproductive health or mental wellness. Among organ-specific question sets, lung had the lowest absence rate (72.5%), suggesting broader topical coverage in lung transplant handbooks, while pancreas had the highest (85.6%).

Reproductive Health had the highest topic-level absence rate (95.1%), indicating an important gap in patient education materials: the vast majority of handbooks do not address fertility, contraception, or pregnancy after transplant. Financial & Administrative topics had the lowest absence rate (72.4%).

3.3 Organ-Level Heterogeneity

Table 5 reports organ-level divergence and consistency rates (Eq. 4). Kidney and lung exhibited the highest divergence: 39.8% and 41.8% of questions had at least one divergent or contradictory pair, respectively. Pancreas showed the lowest divergence prevalence (11.0%), likely reflecting both fewer centers and more standardized guidance for pancreas transplantation.

Table 4: Absence rates by organ type and selected topic categories. The absence rate is the fraction of question–handbook pairs in which the handbook did not address the question (Eq. 1).

Group	Questions	Pairs	Absence rate
<i>By organ type</i>			
General	311	31,722	0.905
Heart	137	3,562	0.818
Kidney	196	4,312	0.780
Liver	164	2,788	0.805
Lung	153	3,978	0.725
Pancreas	154	1,694	0.856
<i>By topic (selected)</i>			
Reproductive Health	291	—	0.951
Transplant Process & Logistics	21	—	0.894
Medications	110	—	0.838
Mental & Emotional Health	107	—	0.829
Monitoring & Follow-up	108	—	0.785
Financial & Administrative	19	—	0.724

Table 5: Organ-level heterogeneity metrics. Q_{total} : total questions for the organ; Q_{active} : questions with ≥ 1 non-absent pair; R_{div} and R_{con} : mean question-level divergence and consistency rates (Eq. 4); %Div: percentage of questions with at least one divergent or contradictory pair.

Organ	Q_{total}	Q_{active}	R_{div}	R_{con}	%Div
General	311	122	0.188	0.285	28.6%
Heart	137	69	0.152	0.186	26.3%
Kidney	196	124	0.237	0.148	39.8%
Liver	164	91	0.209	0.142	30.5%
Lung	153	117	0.152	0.136	41.8%
Pancreas	154	44	0.195	0.223	11.0%

General-type questions had the highest mean consistency rate ($R_{\text{con}} = 0.285$), suggesting that cross-cutting topics such as immunosuppression adherence and lifestyle guidance tend to elicit more uniform recommendations when they are covered. Lung and liver had the lowest consistency rates (0.136 and 0.142), indicating that even when centers address the same organ-specific question, they often differ in the specifics of their recommendations.

3.4 Topic-Level Heterogeneity

Table 6 presents heterogeneity metrics for the 13 topic categories. Monitoring & Follow-up exhibited the highest mean divergence rate ($R_{\text{div}} = 0.277$) with 38.9% of questions showing at least one divergent pair, reflecting well-documented variation in post-transplant surveillance protocols across centers. Lifestyle & Daily Living followed ($R_{\text{div}} = 0.235$, 40.4% divergence prevalence), consistent with the lack of standardized guidelines for diet, exercise, and activity restrictions.

Conversely, Reproductive Health showed the highest consistency rate ($R_{\text{con}} = 0.315$) but the lowest divergence prevalence (17.5%). This initially surprising finding reflects its extreme absence rate (95.1%): the few centers that do address reproductive topics tend to convey similar core messages (e.g., avoid pregnancy in the first year), but the vast majority of handbooks omit this information entirely. Mental & Emotional Health had the lowest divergence ($R_{\text{div}} = 0.087$), suggesting that psychosocial guidance, when provided, is relatively uniform.

3.5 Center-Level Heterogeneity

Among the 23 anonymized centers with sufficient data for analysis, center-level divergence rates ranged from $R_{\text{div}} = 0.139$ to 0.255, and consistency rates from $R_{\text{con}} = 0.082$ to 0.194. The three centers with the highest divergence

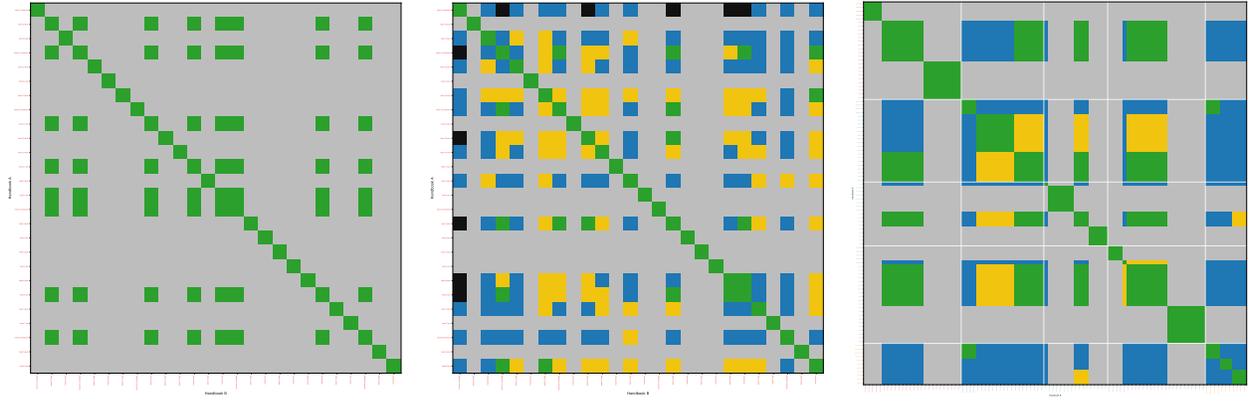
Table 6: Topic-level heterogeneity metrics (topics with ≥ 20 questions). Columns as in Table 5.

Topic	Q_{total}	Q_{active}	R_{div}	R_{con}	%Div
Monitoring & Follow-up	108	59	0.277	0.073	38.9%
Lifestyle & Daily Living	225	134	0.235	0.128	40.4%
Pre-Transplant	176	105	0.221	0.179	29.5%
Medical Complications	319	168	0.179	0.177	26.6%
Reproductive Health	291	85	0.176	0.315	17.5%
Special Populations	73	39	0.172	0.242	24.7%
Surgery & Recovery	78	36	0.154	0.098	29.5%
Financial & Administrative	19	10	0.147	0.216	36.8%
Medications	110	63	0.120	0.233	30.0%
Mental & Emotional Health	107	51	0.087	0.193	19.6%

profiles (center-024, center-019, center-008; $R_{\text{div}} \geq 0.238$) consistently produced answers differing from peer institutions, while the most consistent centers (center-022, center-012; $R_{\text{con}} \geq 0.183$) showed stronger alignment with cross-center norms. This variation across centers was not simply an artifact of sample size: the number of non-absent pairs per center ranged from 764 to 12,323, and divergence patterns persisted after controlling for data volume.

3.6 Illustrative Comparison Matrices

Figure 4 presents three comparison matrices selected to illustrate the range of heterogeneity patterns observed across questions.



(a) Q22 (heart): “Does blood type affect how long you wait for a heart transplant?” All 28 non-absent pairs are CONSISTENT—a clear cross-center consensus.

(b) Q219 (heart): “How should I handle dental care after my transplant?” All five labels present including CONTRADICTION (black), reflecting known clinical controversy over antibiotic prophylaxis.

(c) Q46 (general): “How often should I have blood work done?” $r_{\text{div}} = 0.871$: centers recommend vastly different monitoring schedules.

Figure 4: Selected comparison matrices illustrating heterogeneity patterns. Each cell represents a pairwise comparison between two center handbooks for a single question. Colors: grey = ABSENT, green = CONSISTENT, yellow = COMPLEMENTARY, blue = DIVERGENT, black = CONTRADICTION. Handbook labels on axes are anonymized (organ-center_index-phase). Matrices are symmetric; diagonal entries are CONSISTENT by convention.

These matrices reveal several structural patterns. First, consistency tends to cluster within organ types: in Q22 (Figure 4a), all responding heart centers agree on the role of blood type in waitlist priority, reflecting well-established allocation policy. Second, divergence can emerge even in ostensibly straightforward questions: Q46 (Figure 4c) asks about blood work frequency, yet centers recommend schedules ranging from weekly to every few months, producing $r_{\text{div}} = 0.871$ —the highest in the dataset. Third, the coexistence of all five labels within a single matrix (Figure 4b) demonstrates that institutional disagreement can be highly structured: for dental care after heart transplant, specific

center pairs consistently contradict one another on antibiotic prophylaxis while agreeing with other peers. More broadly, COMPLEMENTARY patterns—in which centers provide medically compatible guidance that differs in scope and emphasis—are the most common non-absent label across the dataset, reflecting institutional preferences in what to highlight for patients.

4 Discussion and Conclusion

4.1 Principal Findings

First, **cross-center divergence is significant but unevenly distributed**. Among 68,019 non-absent pairwise comparisons, 20.8% were classified as DIVERGENT and 0.2% as CONTRADICTORY. Lung and kidney questions showed the highest divergence prevalence (41.8% and 39.8% of questions), while pancreas showed the lowest (11.0%). At the topic level, Monitoring & Follow-up and Lifestyle & Daily Living exhibited the highest divergence rates ($R_{\text{div}} = 0.277$ and 0.235), consistent with the absence of standardized national guidelines for post-transplant surveillance and activity restrictions.

Second, **coverage gaps are a dominant source of information inequality**. The overall absence rate of 96.2% indicates that most handbooks address only a narrow slice of patients' information needs. Reproductive Health had a 95.1% absence rate, meaning the vast majority of handbooks provide no guidance on fertility, contraception, or pregnancy. This systematic omission may be more consequential than outright disagreement, as patients at certain centers receive no information on an important topics rather than merely different information.

Third, **center-level divergence profiles are stable and interpretable**. Divergence rates ranged from 0.139 to 0.255 across centers, suggesting that some institutions systematically depart from cross-center norms in ways that reflect genuine differences in clinical philosophy or authoring conventions rather than sample size artifacts.

Fourth, **individual questions exhibit structured heterogeneity**. The comparison matrices (Figure 4) show that divergence is not random noise: topics such as antibiotic prophylaxis for dental care (Q219) and monitoring frequency (Q46) produce structured patterns of agreement and disagreement reflecting documented clinical controversies. These patterns underscore that document selection in RAG-based medical question answering functions as an implicit clinical decision, as systems grounded in a single center's materials inherit that institution's omissions and positions on contested topics. The coverage gap analysis and center-level profiles can directly support quality improvement by identifying topics where education materials should be expanded and flagging institutions most misaligned with peer consensus. More broadly, the framework is applicable to other domains where multiple institutions issue guidance on overlapping topics, including oncology protocols, chronic disease management, and discharge instructions.

4.2 Limitations

Several limitations should be acknowledged. Although we conducted sample annotation and agreement checks, the LLM-based pairwise judge may still introduce systematic classification biases. The hybrid retrieval pipeline, while effective, is not perfect; more advanced retrieval methods could improve passage selection and reduce noise in the generated answers. Additionally, our pipeline currently processes only the textual content of handbooks, yet many handbooks also contain tables, figures, and infographics that may address questions currently identified as coverage gaps; incorporating multimodal extraction could therefore refine gap estimates. Finally, our corpus is limited to English-language handbooks from 23 U.S. transplant centers, and extending the framework to non-English materials and international transplant systems would improve generalizability.

4.3 Conclusion

This work introduces a scalable framework for quantifying institutional heterogeneity in transplant patient education materials through document-grounded language model comparison. Our analysis of 102 handbooks from 23 U.S. centers reveals that over 20% of non-absent pairwise comparisons reflect clinically meaningful divergence, concentrated in topics like post-transplant monitoring and lifestyle restrictions, while coverage gaps are even more pervasive, with reproductive health exhibiting a 95.1% absence rate. These findings highlight that document selection in retrieval-augmented medical question-answering task functions as an implicit clinical decision, reinforcing current practice. Systems grounded in a single center's materials inherit that institution's omissions and positions on difficult

and contested topics. The topic-level absence and center-level divergence profiles offer actionable benchmarks for harmonizing patient education. By making cross-center variation measurable and structured, our framework provides a foundation for more transparent and comprehensive document-grounded medical question answering to meet the diverse information needs of patients with complex health conditions.

Acknowledgments

This work used Bridges-2 at the Pittsburgh Supercomputing Center (PSC) through allocation CIS250181 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. We sincerely thank all the transplant centers that provided their patient handbooks for this analysis. We are grateful to T. Mace and J. Mace at Transplants.org who assembled and shared the handbooks with us for this study. This research was also funded by the National Institute of Standards and Technology under Federal Award ID Number 60NANB24D231 and Carnegie Mellon University AI Measurement Science and Engineering Center (AIMSEC).

References

1. King KL, Husain SA, Schold JD, Patzer RE, Reese PP, Jin Z, et al. Major variation across local transplant centers in probability of kidney transplant for wait-listed patients. *Journal of the American Society of Nephrology*. 2020;31(12):2900-11.
2. Rivera B, Canizares S, Cojuc-Konigsberg G, Holub O, Nakonechnyi A, Chumdermpadetsuk RR, et al. Examining Transparency in Kidney Transplant Recipient Selection Criteria: Nationwide Cross-Sectional Study. *JMIR AI*. 2025;4:e74066.
3. Rodrigue JR, Feranil M, Lang J, Fleishman A. Readability, content analysis, and racial/ethnic diversity of online living kidney donation information. *Clinical transplantation*. 2017;31(9):e13039.
4. Poudel A, Adhikari A, Poudel S, Poudel A. Readability of online patient education materials related to liver transplantation in the United States. *Transplantology*. 2024;5(3):216-23.
5. Mace T, Mace J, Friedman B, Padman R, Clemente S, Abdulakhadov A, et al. Improving Quality of Patient Educational Materials through a Comparative Analysis of Patient Handbooks from US Transplant Centers. *American Journal of Transplantation*. 2025;25(8):S995.
6. LlamaIndex. LlamaParse; 2024. Document parsing platform for LLM applications. <https://developers.llamaindex.ai/>.
7. Robertson S, Zaragoza H. *The probabilistic relevance framework: BM25 and beyond*. vol. 4. Now Publishers Inc; 2009.
8. Douze M, Guzhva A, Deng C, Johnson J, Szilvasy G, Mazaré PE, et al. The faiss library. *IEEE Transactions on Big Data*. 2025.
9. Xiao S, Liu Z, Zhang P, Muennighoff N, Lian D, Nie JY. C-pack: Packed resources for general chinese embeddings. In: *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*; 2024. p. 641-9.
10. Cormack GV, Clarke CL, Buettcher S. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*; 2009. p. 758-9.
11. Wang W, Wei F, Dong L, Bao H, Yang N, Zhou M. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*. 2020;33:5776-88.
12. Lin J, Ma X, Lin SC, Yang JH, Pradeep R, Nogueira R. Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*; 2021. .