The Cost of Complexity: How Communication Frictions Affect Patient Mortality*

Kevin Kloiber†

November 11, 2025

The most recent version of the paper can be found here.

Abstract

This paper examines how the complexity of medical discharge instructions influences patient mortality. Using data on 239,878 hospital admissions from a large U.S. academic medical center, I measure textual complexity using standard readability metrics. A one-standard-deviation increase in complexity is associated with a 0.14 percentage point rise in 28-day mortality, representing an 8.7% increase relative to the baseline rate. Associations are highly heterogeneous and concentrated among patients requiring intensive self-management: for heart failure patients, the association is nearly ten times larger. The pattern persists within subsamples of notes containing identical self-care instructions, demonstrating that linguistic framing matters beyond task assignment. By leveraging unstructured clinical text, I measure communication quality, an input in the health production function that has remained largely unobservable in administrative data. The magnitude of the mortality association is comparable to estimates from Medicare eligibility or increased hospital spending, suggesting that interventions to simplify discharge instructions could yield substantial health benefits at low cost.

*I would like to express my gratitude to my supervisors, Andreas Peichl and Joachim Winter, for their invaluable guidance and contributions to this paper. I thank Davide Cantoni, Ingrid Hägele, Andrew Proctor, Lars Riedemann, Derya Uysal, Fabian Waldinger, Francis Wong, and Peter Zorn, as well as participants at the EDGE Jamboree 2025, ASHEcon 2025, the HPI-ZEW-DIW Workshop on Applied Economics in Digital Health, and internal seminars in Munich, for helpful comments. This research was supported in part by the Google Cloud Research Credits program. All errors are mine.

†LMU Munich, Department of Economics. E-mail: kevin.kloiber@econ.lmu.de.

"The single biggest problem in communication is the illusion that it has taken place."

— George Bernard Shaw

1 Introduction

A large body of evidence in economics shows that complexity imposes cognitive costs that are first-order determinants of behavior. Simplifying administrative processes generates large behavioral responses: streamlining financial aid applications boosts college enrollment (Bettinger et al., 2012), and clarifying outreach letters for the Earned Income Tax Credit increases program take-up (Bhargava and Manoli, 2015). Experimental evidence confirms that individuals are willing to pay to avoid convoluted rules (Oprea, 2020). In healthcare, prior work has examined how complexity in insurance plan design leads to suboptimal choices and financial losses (e.g., Abaluck and Gruber, 2011; Handel and Kolstad, 2015). Yet these settings involve voluntary choices where individuals can opt out or delay engagement. This paper examines whether complexity in discharge instructions affects mortality following hospital discharge, a setting where complexity cannot be avoided: when patients leave the hospital, they must navigate whatever instructions they receive to manage their recovery, often in a vulnerable state with life-or-death consequences.

Using data on 239,878 hospital admissions from a large U.S. academic medical center, I find that a one-standard-deviation increase in instructional complexity is associated with a 0.14 percentage point increase in 28-day mortality (SE = 0.03), representing an 8.7 percent increase relative to the baseline rate. Moving from the 25th to the 75th percentile of complexity generates a 0.18 percentage point increase in mortality (SE = 0.05). Effects are concentrated where self-care is most intensive: for heart failure patients, the association is nearly ten times larger. I find no statistically significant association with in-system hospital readmissions or emergency department visits. This mortality effect is comparable in magnitude to estimates from Medicare eligibility (Card et al., 2009) or a one-standard-deviation increase in hospital spending (Doyle et al., 2015), yet interventions to improve discharge instruction readability would be substantially less costly to implement.

The analysis focuses on the transition from hospital to home, a period of acute vulnerability during which 1.77 percent of patients in my sample die within 28 days. Discharge instructions mediate this handoff by translating expert knowledge into actionable self-care guidance. Consistent with prior clinical research, I document that for a sick, elderly patient population, average discharge instructions are written at an 11th-grade reading level, far exceeding the 6th-grade level recommended by public health organizations. The data reveal a fundamental disconnect: patients must act on information they may struggle to comprehend. Recent clinical research confirms that poor comprehension of discharge instructions is widespread and associated with lower self-reported adherence (Glick et al., 2019).

This paper makes three contributions. First, it provides the first large-scale evidence linking textual complexity in discharge instructions to mortality. While clinical research documents comprehension struggles (Glick et al., 2019), and prior economics work shows complexity affects financial choices (Abaluck and Gruber, 2011; Handel and Kolstad, 2015), whether readability is associated with survival remains an open empirical question. Second, I develop a model in which linguistic complexity introduces noise into patients' signal extraction problem, generating comprehension errors when translating instructions into self-care actions. The model predicts larger associations where health is highly sensitive to self-care actions. The data confirm this: for heart failure patients requiring intensive daily self-management, the mortality association is nearly ten times larger than average, distinguishing behavioral responses from pure selection. Third, I demonstrate how unstructured clinical text can measure communication quality, a modifiable input in the health production function that has remained largely unobservable to economists. Administrative datasets typically contain observable patient characteristics such as demographics, diagnoses, and procedures. However, many inputs that shape patient behavior fall into an intermediate category: they are observable to physicians and patients at the point of care but unobservable to researchers using standard administrative data. Accessing discharge instructions moves these communication inputs from the unobservable to the observable domain. This methodological contribution allows researchers to assess whether improving information transmission can generate welfare gains comparable to those from medical technology or insurance expansion.

The data come from the Medical Information Mart for Intensive Care (MIMIC-IV), which links detailed clinical information to the full text of discharge summaries from 239,878 hospital admissions between 2008 and 2019 at Beth Israel Deaconess Medical Center, a large tertiary academic medical center in Boston affiliated with Harvard Medical School (Johnson et al., 2023b, 2024, 2023a). I develop a text processing pipeline to extract patient-facing discharge instruction sections and measure complexity using the SMOG (Simple Measure of Gobbledygook) readability index, a validated metric for healthcare settings that maps text to grade-level difficulty (Leonard Grabeel et al., 2018). To explore whether effects vary by instructional content, I employ large language models with structured prompts to identify specific self-care tasks within discharge instructions, such as daily weight monitoring, medication adherence, and dietary restrictions.

A key empirical challenge is that patient severity may correlate with both instructional complexity and mortality. To isolate the relationship, I estimate a high-dimensional linear probability model with rich adjustment for demographics, comorbidities, ICU use, length of stay, and prior utilization, along with fixed effects for clinical diagnosis, provider, service, discharge destination, and year. This approach compares patients in near-identical clinical contexts. Institutional features of the setting provide a source of variation that helps address confounding concerns. At teaching hospitals, inpatient discharge summaries are commonly authored by trainee physicians and then countersigned by attending physicians, with formal training in patient-facing writing remaining limited (Legault et al., 2012; Horwitz et al., 2013; Axon et al., 2014; Black and Colford, 2017). Rotation-driven heterogeneity in trainee authors, limited formal training, and time pressure collectively generate variation in linguistic complexity among patients with similar diagnoses and treatments.

The model predicts larger effects where intensive daily self-management is required. The data confirm this: for heart failure patients, a condition requiring daily weight monitoring, strict dietary adherence, and symptom-based medication adjustments, a one-standard-deviation increase in complexity is associated with a 1.3 percentage point increase in mortality. By contrast, I find no association for conditions where short-term outcomes are dominated by disease progression rather than behavioral inputs. To test whether the effect operates through linguistic framing rather than task content, I examine subsamples of patients who received instructions for identical self-care tasks. The association persists for instructions containing weight monitoring and red-flag symptom warnings, though effects are imprecisely estimated for other task types, likely due to smaller sample sizes. The mortality association is robust across multiple readability metrics, confirming it is not an artifact of the SMOG index specifically.

The temporal pattern supports a behavioral mechanism: the association is statistically insignificant for the first four days post-discharge, emerges by day five, and peaks between weeks three and seven, inconsistent with the hypothesis that complexity merely proxies for imminent mortality. Results remain statistically significant when excluding patients with documented do-not-resuscitate orders, who have substantially higher baseline mortality.

Falsification tests show that complexity is uncorrelated with patient severity at admission, including vital signs and predicted mortality from a machine learning model trained on admission data. The mortality association remains stable when controlling for treatment intensity, including the number of lab tests, procedures, and prescriptions administered during the stay. I verify robustness to alternative functional forms for diagnostic controls and apply sensitivity bounds to quantify how strong selection on unobservables would need to be to eliminate the association.

The null effects on in-system readmissions and emergency department visits warrant careful interpretation. These outcomes likely suffer from classical measurement error, as patients experiencing complications may seek care at other facilities in Boston's multi-hospital region, while mortality is comprehensively captured through Social Security records. Additionally, patients who struggle with complex instructions may actively avoid returning to the discharging hospital, though I cannot directly test this mechanism.

This paper contributes to the literature on how informational frictions influence economic decisions and health outcomes. Theoretical work on bounded rationality and costly attention shows that complexity can distort choices (Gabaix, 2014; Caplin and Dean, 2015). Empirical studies demonstrate that simplifying administrative processes generates large behavioral responses in domains ranging from college enrollment (Bettinger et al., 2012) to public program participation (Bhargava and Manoli, 2015) and retirement saving (Madrian and Shea, 2001). In healthcare, prior work has focused on how complexity in insurance plan design leads to financial losses (Abaluck and Gruber, 2011; Heiss et al., 2013; Ho et al., 2017). This paper shifts the focus from ex-ante plan choice to ex-post adherence, showing that complexity in discharge instructions is associated with mortality.

The paper also contributes to understanding what inputs produce health. While foundational work documents the roles of medical technology and insurance (Cutler and McClellan, 2001; Finkelstein, 2007), and recent research emphasizes provider heterogeneity (Chandra and Staiger, 2007; Doyle et al., 2015), communication quality remains an overlooked input. The findings show that communication clarity is a modifiable factor with direct effects on survival.

Beyond general complexity, provider-patient communication is a core issue in the economics of agency. Since Arrow (1963), economists have understood medicine as exemplifying an information asymmetry in which patients rely on providers for knowledge. Discharge instructions serve as a communication device intended to mitigate this agency problem and enable patients to co-produce their own health. The literature in medicine documents persistent readability gaps in patient-facing materials and associations with self-reported comprehension (Glick et al., 2019), but prior quantitative studies linking readability to hard outcomes have been limited. Existing work either examines small samples using qualitative methods or focuses on different outcomes, such as readmissions (Russell et al., 2024; Burns et al., 2022). This paper provides the first large-scale evidence linking textual complexity to mortality, showing that the association is concentrated among patients requiring intensive self-management, which distinguishes this contribution from health services research on health literacy.

Methodologically, this paper contributes to the growing use of unstructured text in economics. While economists have increasingly leveraged text to measure beliefs, attention, and persuasion (Gentzkow and Shapiro, 2010; Gentzkow et al., 2019), this paper applies these tools to measure communication quality in a high-stakes health setting and demonstrates how large language models can categorize clinical text at scale. This enables tests of whether associations vary by instructional content, moving beyond broad diagnostic categories to probe heterogeneity based on the specific textual content of provider-patient interactions.

The remainder of this paper proceeds as follows. Section 2 develops the conceptual framework and provides background information. Section 3 details the data and measurement. Section 4 outlines the empirical strategy. Section 5 presents results. Section 6 discusses policy implications, and Section 7 provides a conclusion.

2 Background and Conceptual Framework

Establishing whether instructional complexity affects mortality requires addressing a fundamental identification problem: note complexity is not randomly assigned. Physicians write more complex instructions for certain patients, and those characteristics may independently determine survival. I develop a stylized model to clarify what variation I exploit and generate testable predictions. The framework predicts that harm from complexity concentrates where health is highly sensitive to self-care actions: patients whose survival depends on daily behavioral inputs (high φ) suffer disproportionately from comprehension errors. I use this prediction to test whether observed associations reflect behavioral responses rather than pure selection.

The model motivates the specification and predicts where heterogeneity should emerge. I do not attempt causal identification through quasi-experimental variation, nor estimate structural parameters. Before presenting the framework, I describe how discharge instructions are authored and delivered.

Background. Hospital discharge summaries serve two functions. The bulk is a technical summary for the next physician. Embedded within is a patient-facing section with discharge instructions—the focus of my analysis. This section translates medical care into actionable guidance: medication management, diet and activity restrictions, symptom monitoring, and follow-up appointments.

While the discharge summary and its patient-facing instructions are written by physicians or physician trainees, the delivery of these instructions at discharge typically involves nursing staff. Nurses are often the last providers to interact face-to-face with patients and review the written instructions verbally before the patient leaves. This dual-channel (written + verbal) communication is important for interpretation: the linguistic complexity of the written document reflects the physician's writing choices, but patients' actual comprehension may be mediated by nurses' verbal explanations. My data contain only the written discharge note text, not information about the verbal discharge process or which nurses reviewed instructions with which patients. The measured associations, therefore, reflect exposure to written complexity, which may understate total communication quality if nurses effectively translate complex instructions orally, or overstate it if nurses reinforce confusing text.

Providing patient-facing discharge instructions is not merely a clinical best practice; it is a federal mandate. Under the Centers for Medicare & Medicaid Services (CMS) Conditions of Participation, hospitals are legally required to implement a comprehensive discharge planning process that provides patients with understandable information to ensure a safe transition to post-hospital care.[1] Failure to meet these conditions can jeopardize a hospital's certification to receive Medicare payments, though in practice hospitals typically remedy deficiencies before actual decertification occurs. Layered on top of this mandate are direct financial incentives. Since 2012, the Hospital Readmissions Reduction Program (HRRP) has penalized hospitals for excess readmissions for conditions like Heart Failure and COPD.[2] Clinical evidence provides the rationale for this policy; systematic reviews demonstrate that communication interventions at hospital discharge are associated with lower readmissions (Becker et al., 2021; Kripalani et al., 2014; Hesselink et al., 2014). The HRRP therefore creates a strong financial motive for hospitals to ensure their discharge communication is effective.

Despite these mandates and incentives, a fundamental disconnect persists. In my sample of 239,878 hospital admissions, the average discharge note is written at an 11th-grade reading level, far exceeding the 6th-grade level recommended by public health organizations for patient-facing materials. For a patient population that is sick, elderly, and cognitively strained by recent hospitalization, this readability gap represents a potentially severe informational friction. The question is whether this friction has measurable consequences for survival.

Understanding this disconnect requires examining how discharge instructions are produced. At BIDMC, a large academic teaching hospital, inpatient discharge summaries are commonly authored by trainee physicians (interns and residents) and then countersigned by attending physicians (Legault et al., 2012; Axon et al., 2014; Horwitz et al., 2013). Residents are responsible for authoring discharge summaries for patients admitted to their team, yet few receive formal training in writing patient-facing documents (Black and Colford, 2017; Ming et al., 2019). Throughout a patient's stay, information accumulates in the Electronic Health Record across multiple providers. Generating the discharge summary requires manually assembling this information, with physicians spending considerable time writing the narrative hospital course. Following completion, the discharge summary is forwarded to the attending physician for signature, a process that often takes several days. Surgical notes are typically written by residents and signed by attendings, further decoupling operative skill from note readability (You et al., 2022). The constant rotation of a large pool of trainees with heterogeneous writing styles, educational backgrounds, and varying propensity

[1] 42 CFR § 482.43, Condition of Participation: Discharge planning.

[2] Section 3025 of the Patient Protection and Affordable Care Act of 2010.

to use EHR templates generates considerable variation in the complexity of instructions received by clinically similar patients.

A Model of Costly Cognition. I develop a stylized model that formalizes how linguistic complexity affects health outcomes through imperfect comprehension of self-care instructions. The framework builds on the health production function literature (Grossman, 1972) and incorporates information frictions in the spirit of recent work on costly attention in economic decision-making (Gabaix, 2014; Caplin and Dean, 2015).

Setup. The patient chooses a self-care action $a \in \mathbb{R}$ to maximize expected health. Health is produced according to:

$$h = \varphi \cdot a \tag{1}$$

where $\varphi > 0$ captures the sensitivity of health outcomes to self-care actions. The parameter φ varies across patients and clinical conditions: it is high for conditions where intensive daily self-management directly affects near-term outcomes (e.g., Heart Failure, where daily weight monitoring and medication titration prevent decompensation) and low for conditions where outcomes are dominated by disease biology rather than behavioral inputs (e.g., sepsis, cancer).

There exists an optimal action a^* that would maximize the patient's health. However, the patient does not directly observe a^*. Instead, they must extract it from the discharge instructions, which provide a noisy signal:

$$s = a^* + \varepsilon$$

where $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ represents comprehension error. This is a standard signal extraction problem with exogenous information structure.

The variance of the comprehension error depends on the note's characteristics: its linguistic complexity, C, proxied empirically by the SMOG readability score, and its information content, I, proxied by the log of the note's character count:

$$\sigma_\varepsilon^2 = \psi(C, I)$$

where $\psi_C > 0$ (greater complexity increases comprehension error), $\psi_I < 0$ (more information reduces error through redundancy), and $\psi_{CC} \geq 0$ (complexity has increasing marginal effects on error). The cognitive cost function formalizes that complex language imposes processing costs, requiring patients to allocate scarce cognitive resources to linguistic decoding rather than substantive understanding (Gabaix, 2014; Caplin and Dean, 2015).

Patient's Problem. The patient chooses action a to maximize expected health utility with quadratic loss:

$$\max_{a} \quad E\left[-(h-h^*)^2 \mid s\right] \tag{2}$$

where $h^* = \varphi \cdot a^*$ is optimal health. Given the noisy signal s, the patient's optimal action is $\hat{a} = E[a^*|s] = s = a^* + \varepsilon$. This yields:

$$E\left[-(h-h^*)^2\right] = E\left[-(\varphi \cdot \varepsilon)^2\right] = -\varphi^2 \cdot \sigma_\varepsilon^2 = -\varphi^2 \cdot \psi(C, I)$$

This expression clarifies three channels through which note characteristics affect health. First, holding I fixed, higher complexity C increases $\psi(C, I)$, reducing expected health. Second, holding C fixed, greater information content I reduces $\psi(C, I)$, improving health. Third, the squared term φ^2 amplifies both effects: for patients where actions matter more for health outcomes, the stakes of comprehension errors are larger.

Testable Predictions. The model yields several predictions that guide the empirical analysis.

(1) *Main effects.* Taking derivatives:

$$\frac{\partial E[-(h-h^*)^2]}{\partial C} = -\varphi^2 \cdot \psi_C(C, I) < 0$$
$$\frac{\partial E[-(h-h^*)^2]}{\partial I} = -\varphi^2 \cdot \psi_I(C, I) > 0$$

Holding information content fixed, higher complexity reduces expected health. Holding complexity fixed, more information improves expected health. In the empirical analysis, I proxy complexity with the SMOG readability score and information content with log note length. The model therefore predicts that conditional on a rich set of patient and clinical controls, higher SMOG scores should be associated with higher mortality, while longer notes should be protective.

(2) *Interaction between complexity and information content.*

$$\frac{\partial^2 E[-(h-h^*)^2]}{\partial C \partial I} = -\varphi^2 \cdot \psi_{CI}(C,I)$$

If $\psi_{CI} < 0$, additional information partially offsets the harm of complexity. Empirically, this motivates controlling for note length when estimating the association between complexity and mortality.

(3) *Heterogeneity by self-care sensitivity.* The central prediction is that harm from complexity concentrates where self-care matters most. Taking the cross-derivative with respect to φ:

$$\frac{\partial^2 E[-(h-h^*)^2]}{\partial C \partial \varphi} = -2\varphi \cdot \psi_C(C,I) < 0$$
$$\frac{\partial^2 E[-(h-h^*)^2]}{\partial I \partial \varphi} = -2\varphi \cdot \psi_I(C,I) > 0$$

The marginal harm of complexity and the marginal benefit of information are both increasing in φ. Patients whose health is highly sensitive to self-care actions suffer disproportionately from linguistic barriers and benefit more from additional information.

To test this prediction, I link the theoretical parameter φ to observable patient characteristics. I operationalize high φ through clinical conditions that require intensive, daily self-management where adherence directly affects near-term outcomes. Heart Failure and COPD exhibit high φ: patients must track daily weights, interpret fluctuating symptoms, adjust medications, adhere to dietary restrictions, and recognize warning signs (Heidenreich et al., 2022; Greene et al., 2015; Agustí et al., 2023). These conditions exhibit high φ precisely because daily behavioral inputs (medication adherence, symptom monitoring, dietary compliance) directly translate into near-term health through rapid physiological mechanisms. By contrast, sepsis and cancer exhibit low φ for the 28-day horizon: near-term mortality is dominated by disease severity rather than day-to-day behavioral inputs (Courtright et al., 2020; Engoren et al., 2023).

The model predicts that the association between note complexity and mortality should be substantially larger for Heart Failure and COPD patients than for sepsis and cancer patients. This heterogeneity test distinguishes between two interpretations of the baseline association. If complexity merely proxies for unobserved patient severity—with sicker patients both receiving more complex notes and independently having higher mortality—then the association should be similar across patient types once I control for observables. By contrast, if complexity operates through comprehension barriers as the model predicts, effects should concentrate precisely where self-care sensitivity is high. Finding larger associations for Heart Failure and COPD than for sepsis and cancer would support a behavioral interpretation over pure selection on severity.

Identification Challenge. The complexity (C) and information content (I) of the note are endogenous, chosen by physicians based on observable patient characteristics X_i (demographics, diagnoses, comorbidities) and potentially the health production parameter φ_i. This creates confounding: if X_i affects both note complexity and mortality directly, estimates of complexity's effect will be biased. My empirical strategy addresses this by including extensive controls and fixed effects that absorb variation in X_i: demographics, comorbidities, ICU use, length of stay, prior utilization, and fixed effects for Diagnosis-Related Groups, the patient's top three diagnosis codes, admitting provider, hospital service, discharge destination, and admission year. This approach compares patients in near-identical clinical contexts. I then test for residual confounding through falsification checks, temporal analysis, and sensitivity bounds in Section 5.2.

3 Data and Measurement

This section describes the data sources, sample construction, and measurement of discharge instruction complexity. I draw on comprehensive electronic health records from a large academic medical center, linking clinical data to the full text of discharge summaries. The primary challenge is measuring textual complexity in a way that captures cognitive demands on patients while remaining comparable across notes of varying length and clinical content.

Data. This study combines two linked datasets: the Medical Information Mart for Intensive Care (MIMIC-IV, v3.1), a comprehensive, de-identified EHR database from the Beth Israel Deaconess Medical Center (BIDMC) covering admissions from 2008–2019 (Johnson et al., 2023b, 2024; Goldberger et al., 2000), and the MIMIC-IV-Note database, which contains the deidentified free-text clinical notes for these admissions (Johnson et al., 2023a). The unit of analysis is a hospital admission. To construct the analysis sample, I first isolate index hospital admissions by excluding any admission that was preceded by another hospital stay for the same patient within the prior 28 days. This step ensures the sample consists of initial hospitalizations rather than early readmissions. From this set of index admissions, I apply further exclusion criteria, removing cases where the patient died in the hospital, was discharged to hospice care, or left against medical advice. Finally, to ensure the quality of the primary explanatory variable, I exclude admissions with missing or exceptionally short discharge notes.

Appendix Table A1 details the sample construction. The excluded group (N 22,800) has a 28-day mortality rate of 9.2%, largely driven by approximately 3,000 patients discharged to hospice care. Excluding hospice patients, the remaining excluded group (those with short, missing, or outlier notes) has a 28-day mortality rate of 1.64%, which is comparable to the 1.77% rate in the final analysis sample. This pattern indicates that note-based exclusions do not induce selection on mortality risk. The final sample comprises 239,878 admissions.

The primary outcome is all-cause 28-day mortality, obtained from linked Social Security Administration death records, which provides a reliable, comprehensive measure of mortality. Short-term mortality measures such as 28-day and 30-day mortality are standard outcomes in health economics research on hospital quality and effectiveness (Chan et al., 2023; Doyle et al., 2015, 2019). Secondary outcomes are in-system 28-day readmissions and emergency department (ED) visits; the in-system nature of these measures means they are a lower bound on total post-discharge utilization. The primary explanatory variable is the linguistic complexity of the discharge instructions, measured by the SMOG index. As dictated by the conceptual framework, all specifications control for information content using the log of the character count of the instructions. Table 1 presents summary statistics. The data reveal the core tension of the paper: a sick, elderly patient population (mean age 62, mean Charlson Index 3.9, indicating high comorbidity burden) receives instructions written at a mean 11th-grade reading level.

Measurement. The primary explanatory variable, note complexity, is measured using the SMOG index (Mc Laughlin, 1969), a validated readability metric for healthcare settings (Leonard Grabeel et al., 2018; Wallace and Lennon, 2004). SMOG (Simple Measure of Gobbledygook) maps text to U.S. grade-level equivalents: a score of 11 corresponds to the reading level of a high school junior, while a score of 6 corresponds to the reading level of early middle school. The index is calculated by

Table 1: Descriptive Statistics for the Full Analysis Sample

	Mean (SD) or Share		Mean (SD) or Share
Outcomes		**Admission Characteristics**	
28-Day Mortality	0.018 (0.127)	Length of Stay (Days)	5.131 (6.642)
28-Day Readmission	0.171 (0.376)	ICU Admission	0.187 (0.390)
28-Day ED Revisit	0.166 (0.372)	Discharged to Home	0.426 (0.495)
7-Day Mortality	0.004 (0.062)	Discharge Hour (0–23)	15.3 (2.5)
7-Day Readmission	0.068 (0.251)	**Admission Source**	
7-Day ED Revisit	0.073 (0.259)	Source: (% Emergency Room)	53.0
Note Characteristics		Source: (% Physician Referral)	25.2
SMOG Readability (Grade Level)	11.14 (1.41)	Source: (% Transfer from Hospital)	10.8
Log Note Length	6.79 (0.64)	Type: (% Emergency Ward)	42.7
Patient Characteristics		Type: (% Observation Admit)	16.0
Age	61.8 (18.2)	Type: (% EU Observation)	10.4
(% Female)	51.3	Type: (% Surgical Same Day)	10.2
Charlson Comorbidity Index	3.9 (2.9)		
Insurance		**Prior Utilization (1 Year)**	
(% Medicare)	49.4	Admissions in Prior Year	0.82 (1.70)
(% Medicaid)	15.9	ED Visits in Prior Year	0.90 (2.13)
Race			
(% White)	68.8	**Clinical Service (Most Frequent)**	
(% Black)	14.8	(% MED)	44.1
Marital Status		(% CMED)	10.3
(% Married)	43.8	(% SURG)	9.8
(% Single)	34.1		
Language			
(% English)	90.0		
Observations (Admissions)	239,878	**Unique Patients**	131,223

Notes: The table presents summary statistics for the analysis sample of 239,878 hospital admissions. Continuous and binary variables are reported as Mean (SD); categorical variables are reported as shares of the full sample. Insurance and race headers list the selected levels; remaining levels are grouped as 'Other/Unknown' where applicable. Prior utilization is measured over the 365 days preceding the index admission.

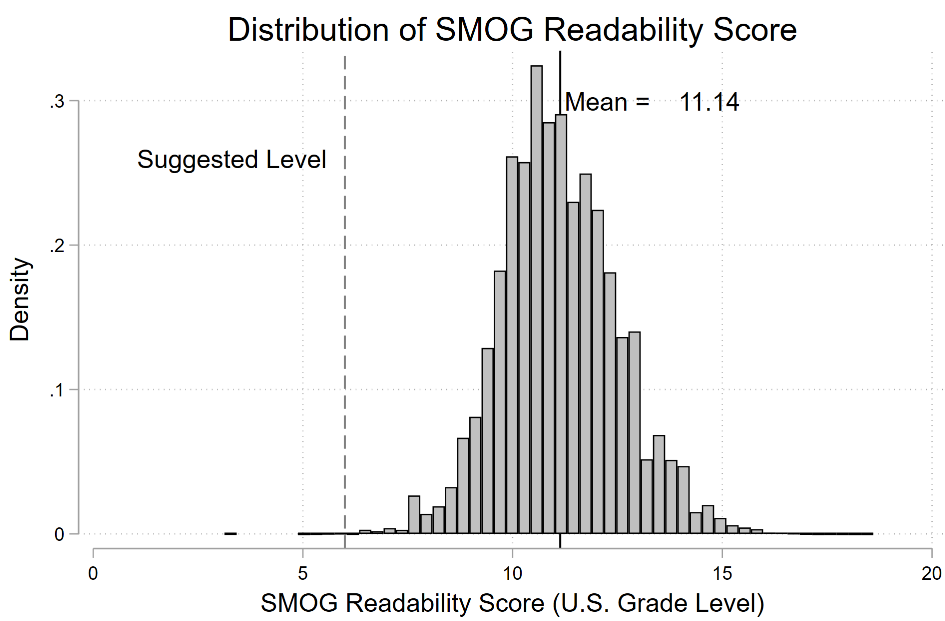
counting polysyllabic words (those with three or more syllables) per sentence. Higher polysyllabic density increases the grade level because longer words impose greater decoding demands on readers. For example, a note scoring at the 6th-grade level might use phrases like "follow a low salt diet," while an 11th-grade note might say "adhere to a low-sodium regimen," reflecting the difference in lexical complexity even when conveying similar clinical content.

Because the SMOG formula is sensitive to sentence counts, a critical step is to reliably identify sentence boundaries in clinical notes that often contain lists, fragments, and inconsistent punctuation. To do this, I developed a text-processing pipeline that first extracts the patient-facing instruction block, then normalizes formatting, converts lists into prose, and segments the text into analyzable sentences.[3] The full technical details are provided in Appendix A.2. I demonstrate that the

[3] Results are similar using scispaCy, a natural language processing toolkit designed for biomedical and clinical text;

main findings are robust to the use of five alternative readability measures: Flesch-Kincaid Grade Level, Gunning Fog Index, Dale-Chall, Coleman-Liau Index, and a masked version of SMOG that replaces medication names. These represent the most widely used readability formulas in health literacy research (Wang et al., 2013), with the Flesch-Kincaid and SMOG measures being the two most commonly applied in healthcare settings. Figure 1 shows the resulting distribution: the mean readability is at an 11th-grade level, well above the 6th-grade target commonly recommended for patient materials.

Figure 1: Distribution of Discharge Instruction Readability (SMOG)



Notes: The figure shows the distribution of SMOG readability scores for patient-facing discharge instructions (N = 239,878 admissions). The dashed vertical line marks the 6th-grade reading level commonly recommended for patient-facing health materials.

While SMOG is the primary measure throughout the analysis, concerns about whether it captures task complexity versus linguistic style motivate an alternative measure. SMOG indexes lexical difficulty based on polysyllabic word density, rising with medical terminology (e.g., ”hypertension,” ”anticoagulation”), Latinate vocabulary, and long sentences, regardless of whether a sentence is an instruction or a narrative background.

The alternative measure, the Clinical Instruction Complexity Score (CICS), is constructed using

I retain a lightweight rule-based approach for computational parsimony.

structured large language model extraction (Gemini via Vertex AI; full details in Appendix A.3). CICS captures procedural complexity within instructions, which involves determining how many contingencies, thresholds, exceptions, and conditional branches a patient must process. Unlike SMOG, CICS does not depend on word length but on branching logic. Figure B1 illustrates this distinction with paraphrased examples from the corpus. The two measures are only weakly correlated (Appendix Table A5), allowing me to examine whether associations persist when comparing notes with similar task content but different lexical complexity. Results from specifications that control for CICS are reported in Appendix B11.

4 Empirical Strategy

I estimate high-dimensional linear probability models to document the conditional association between discharge instruction complexity and post-discharge mortality, controlling for patient characteristics, clinical factors, and institutional features. The goal is not to identify the causal effect of complexity through quasi-experimental variation, but rather to document whether the association persists after adjusting for observable confounders that could spuriously link note complexity to patient outcomes. I begin by describing the main specification, then discuss the source of identifying variation, and finally outline tests for heterogeneous effects motivated by the conceptual framework.

4.1 Main Specification

The primary analysis estimates the conditional association using a linear probability model:

$$y_i = \beta\,\mathrm{SMOG}_i + \gamma\ln(\mathrm{length}_i) + \mathbf{X}_i'\mathbf{\Gamma} + \mathbf{\Lambda_i} + \varepsilon_i, \tag{3}$$

where y_i is a 28-day outcome for hospital admission i. The key regressor is SMOG_i, the grade-level readability score of the discharge instructions. Following the conceptual framework, I control for note information content using the log character count, $\ln(\mathrm{length}_i)$.

The vector $\mathbf{X}_i$ includes patient demographics (age, gender), clinical severity measures (Charlson Comorbidity Index, ICU admission, length of stay), and prior healthcare utilization (admissions and ED visits in the prior year). The vector $\mathbf{\Lambda_i}$ contains fixed effects that absorb variation across patient demographics (language, race, insurance), clinical pathways (top three five-digit ICD codes, DRG, hospital service), and logistical features (admission and discharge locations, admission type, discharge hour, admission year, admitting provider). A detailed breakdown of the number of levels for each fixed effect is provided in Appendix Table B1.

This specification compares patients who are clinically similar (same diagnoses, DRG, service line), logistically similar (same discharge destination, admission pathway), and treated by the same admitting provider, but who receive discharge instructions that differ in linguistic complexity. Standard errors are clustered at the patient level to account for multiple admissions by the same individual.

4.2 Source of Variation in Note Complexity

The variation in note complexity that remains after controlling for the extensive set of covariates described above reflects primarily the discretionary writing choices of the clinician who authors the discharge note. At BIDMC, discharge summaries are typically written by rotating physician trainees (interns and residents) rather than attending physicians. These trainees have diverse writing styles, educational backgrounds, and varying propensities for using EHR templates. Formal training in writing plain-language patient materials is limited (Black and Colford, 2017; Ming et al., 2019). Generating the discharge summary requires manually assembling information that has accumulated across multiple providers during the hospital stay, often under time pressure. Surgical notes are typically written by residents and signed by attendings, further decoupling operative skill from note readability (You et al., 2022).

This institutional context generates two features relevant for interpreting the conditional association. First, the constant rotation of trainees results in substantial variation in the complexity of instructions received by patients with similar clinical conditions. Second, the workflow often decouples the primary clinical expert (e.g., the surgeon who performed a procedure) from the author of the patient-facing discharge instructions, weakening any mechanical relationship between unobserved case complexity and note readability. Note that the admitting provider fixed effects in Equation (3) capture the admitting clinician, not the author of the discharge note, and therefore do not absorb variation in writing style.

While this institutional feature strengthens the plausibility of the research design, it does not guarantee that complexity is uncorrelated with unobserved determinants of outcomes. Factors such as heightened attending oversight for the sickest patients or unobserved resident skill could still introduce a correlation between note complexity and mortality risk. I therefore interpret β as a conditional association rather than a causal effect, and I explicitly test for residual confounding using falsification checks in Section 5.

4.3 Heterogeneity by Self-Care Sensitivity

The conceptual framework predicts that the association between complexity and mortality will be concentrated among patients whose conditions exhibit high self-care sensitivity (high φ), specifically those for whom intensive daily self-management directly affects near-term health outcomes. To test this prediction, I estimate Equation (3) separately on pre-specified patient subsamples defined by clinical condition.

The high-φ subsample includes patients with Heart Failure or COPD. These conditions are characterized by intensive daily self-care requirements. Heart Failure patients must monitor their daily weights, interpret fluctuating symptoms (e.g., increased swelling, shortness of breath), titrate diuretics based on symptoms, adhere to strict dietary sodium limits, and recognize warning signs that require urgent action. COPD patients face similar demands around symptom monitoring, inhaler technique, and exacerbation recognition. These conditions exhibit high φ because daily behavioral inputs (medication adherence, symptom monitoring, dietary compliance) directly translate into near-term health through rapid physiological mechanisms. Missed doses of diuretics can lead to fluid reaccumulation within days, while failure to recognize warning symptoms delays necessary care-seeking. I define patients as having Heart Failure or COPD if any diagnosis code from their current admission or documented medical history includes ICD-10 codes I50.* (Heart Failure) or J44.* (COPD), or their ICD-9 equivalents (428.*, 491.*, 492.*, 496.*).

For comparison, I estimate the same specification for patients with sepsis and cancer. While these conditions carry comparable baseline mortality risk, they exhibit low φ for the 28-day horizon: near-term outcomes are dominated by disease severity, organ dysfunction resolution, and treatment response rather than day-to-day behavioral inputs. Short-term survival after sepsis primarily depends on the resolution of the acute illness and underlying organ function. Similarly, for cancer patients, the short-term disease trajectory is largely determined by cancer progression and treatment toxicity. This makes them a valid comparison group for testing whether complexity associations concentrate where self-care sensitivity is high, as the model predicts.

4.4 Exploratory Analysis by Instructional Content

To probe whether associations reflect general lexical difficulty or specific task contexts, I leverage large language model extraction to identify subsamples of notes containing common self-care instructions (e.g., weight monitoring, medication changes). The extraction method is detailed in Appendix A.3. I re-estimate Equation (3) within these content-defined subsamples. This analysis does not represent a clean test of framing versus content, as the presence of specific instructions is endogenous to patient condition. Rather, it assesses whether the association persists within more homogeneous action spaces.

5 Results

This section documents the association between the complexity of discharge instructions and patient outcomes in three steps. I begin by estimating the baseline association using the specification described in Section 4. I then assess whether the association reflects a genuine relationship or spurious correlation through six robustness checks. I show that complexity does not predict admission severity, that effects emerge gradually rather than immediately, and that results persist after excluding palliative patients. I then verify robustness to the addition of explicit controls for treatment intensity and to allowing full interactions between diagnoses. Finally, I apply the method of Oster (2019) to bound the influence of remaining unobservables.

Having established the robustness of the baseline association, I test for heterogeneous effects. I examine whether the association is concentrated in patient cohorts requiring intensive self-management, and then explore whether it persists when holding task content fixed by analyzing subsamples defined by specific instructional content identified using large language models.

5.1 Baseline Association

Table 2 presents estimates from high-dimensional linear probability models from equation 3. Columns (1) through (4) show how coefficients on note complexity (SMOG) and information content (log note length) evolve as I add sequential control blocks. Panel A reports results for 28-day mortality. In the preferred specification (Column 4), a one-grade-level increase in SMOG is associated with a 0.10 percentage point increase in mortality (SE = 0.03). A one-standard-deviation increase (1.4 grade levels) corresponds to a 0.14 percentage point increase, or 7.9 percent relative to the baseline rate of 1.77 percent. Moving from the 25th to 75th percentile of complexity is associated with a 0.18 percentage point increase (SE = 0.05), representing 10.2 percent of the baseline rate.

The association between complexity and adverse outcomes is specific to mortality. Panels B and C show that the coefficient on SMOG is economically small and statistically insignificant for both in-system readmissions and ED visits. These null effects warrant interpretation. First, the dependent variables likely suffer from classical measurement error. BIDMC accounted for only 8.2 percent of adult ED visits among trauma centers in its regional emergency medical services

Table 2: Note Characteristics and Post-Discharge Outcomes

	(1)	(2)	(3)	(4)
Panel A: 28-Day Mortality (Mean = 1.77)				
SMOG	0.02	0.15***	0.11***	0.10***
	(0.02)	(0.02)	(0.03)	(0.03)
Log Note Length	−0.28***	−0.47***	−0.47***	−0.35***
	(0.04)	(0.04)	(0.06)	(0.07)
Panel B: 28-Day Readmission (Mean = 17.96)				
SMOG	−0.09*	−0.01	−0.10	−0.03
	(0.06)	(0.05)	(0.07)	(0.08)
Log Note Length	−0.26**	0.12	0.47***	0.40**
	(0.12)	(0.12)	(0.18)	(0.18)
Panel C: 28-Day ED Visit (Mean = 17.25)				
SMOG	−0.11*	−0.06	−0.08	−0.02
	(0.05)	(0.05)	(0.07)	(0.08)
Log Note Length	0.66***	1.46***	0.54***	0.44**
	(0.12)	(0.12)	(0.18)	(0.18)
Observations	239,878	236,720	190,305	189,859
Controls	Basic	+ Demographics	+ Clinical	+ Institutional

Notes: Linear probability model estimates. Dependent variables listed in panel headers. Coefficients multiplied by 100 (percentage points). Column (1): SMOG and log note length only. Column (2): adds patient demographics. Column (3): adds clinical severity and diagnosis fixed effects. Column (4): adds institutional factors and provider fixed effects (preferred specification). Standard errors clustered at patient level in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

catchment area in fiscal year 2014.[4] I observe only in-system utilization; events at other facilities are unobserved. Patients experiencing post-discharge complications may seek or be transported to other facilities, particularly via ambulance, which quasi-randomly assigns patients across hospitals based on location and capacity (Doyle et al., 2015). This measurement error attenuates any true association between note complexity and total utilization toward zero. Second, patients who struggle with complex instructions may avoid returning to the discharging hospital, seeking care elsewhere or delaying care. While I cannot test this mechanism directly, it represents a plausible channel through which communication barriers affect care-seeking without manifesting in observed in-system utilization.

The coefficient on log note length highlights an identification challenge. Note length can proxy for beneficial information or for unobserved case complexity. For mortality (Panel A), the coefficient is negative and statistically significant across all specifications, consistent with the information provision hypothesis. For readmissions and ED visits (Panels B and C), the coefficient is positive and significant in the preferred specification, suggesting that note length primarily captures severity for these outcomes.

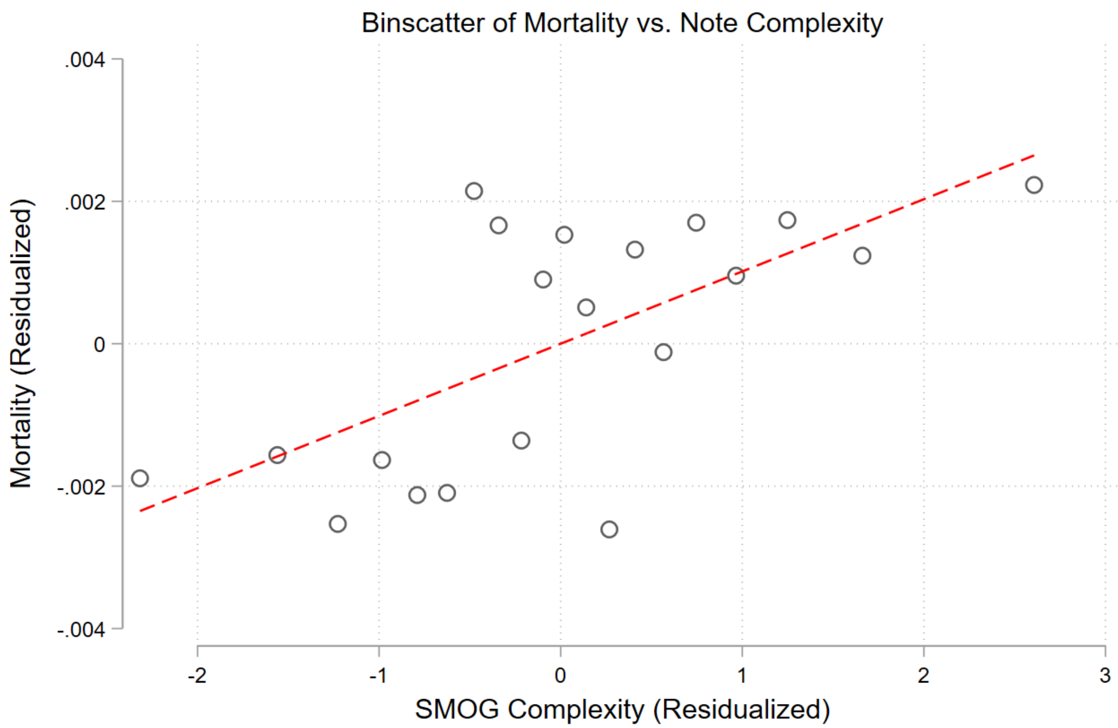
Figure 2 plots the partial relationship between 28-day mortality and SMOG. Both variables are residualized on log note length, covariates, and fixed effects. The slope corresponds to the coefficient in Column (4). The relationship is positive and approximately linear, with no evidence of strong non-linearities.

5.2 Addressing Potential Confounders

The conditional association documented in Table 2 could reflect spurious correlation through several channels. First, complexity may proxy for end-of-life status or baseline frailty: if sicker or palliative patients systematically receive more complex notes, the mortality association would be spurious and should appear immediately rather than emerging gradually over time. Second, complexity may be associated with treatment intensity: patients who undergo more procedures and medications have longer, more complex notes and an independently higher mortality risk. Third, additive diagnosis fixed effects may inadequately control for comorbidity interactions. I address these threats through

[4] Author's calculation from hospital profiles published by the Massachusetts Center for Health Information and Analysis.

Figure 2: Partial Relationship Between 28-Day Mortality and SMOG Complexity



Notes: Both variables are residualized on log note length, the full covariate set, and fixed effects from the preferred specification. By the Frisch-Waugh-Lovell theorem, the slope corresponds to the conditional coefficient in Column (4) of Table 2. The binned scatter plot shows 20 equal-sized bins.

six tests. I show that complexity does not correlate with severity measures fixed at admission, that the mortality association emerges gradually over weeks rather than immediately, and that results persist after excluding patients with Do Not Resuscitate orders. I then control for treatment intensity and verify robustness to fully interacted diagnosis combinations. Finally, I apply the method of Oster (2019) to bound the influence of remaining unobservables.

I first test whether note complexity correlates with patient severity at admission. If physicians write more complex notes for sicker patients, the mortality association would reflect baseline frailty rather than failed self-care. Discharge note complexity, determined at discharge, should not predict characteristics fixed at admission if it reflects discretionary writing choices. Table 3 regresses admission-time measures on SMOG score, conditional on the full control set and fixed effects.

Columns (1) through (3) examine clinical measures: triage acuity (zero for non-ED admissions),

Table 3: Falsification Tests: Note Complexity and Admission Severity

	(1) Acuity	(2) Heart Rate	(3) Resp. Rate	(4) Predicted Mort.
SMOG	-0.19	-2.8	-1.7	-0.002
	(0.17)	(2.5)	(1.1)	(0.004)
Outcome mean	2.15	83.4	19.0	1.54
Observations	189,859	77,403	76,463	189,859

Notes: Each column regresses an admission-time measure on discharge note SMOG score. Coefficients multiplied by 100. All specifications include the full control set and fixed effects from Column (4) of Table 2. Column (4) uses predicted mortality from a LASSO logit trained on admission covariates (AUC = 0.829 on held-out sample). Standard errors clustered at the patient level in parentheses.

heart rate, and respiratory rate (both measured within 24 hours; missing values reduce sample size). Column (4) uses predicted 28-day mortality from a LASSO logit trained on 70 percent of the data using admission covariates (demographics, vital signs, comorbidities, admission characteristics), with missing vitals imputed at sample means. The model achieves an AUC of 0.829 on the held-out sample. Across all measures, the SMOG coefficient is economically negligible and statistically insignificant. Results are robust to including a quadratic term for SMOG in columns (2) and (3); both linear and quadratic coefficients remain insignificant.

A second test examines the temporal evolution of the association to distinguish between competing mechanisms. If complexity merely proxies for unobservable frailty or imminent end-of-life status, the mortality effect should appear immediately at discharge. If, instead, complexity operates through failed adherence, effects should emerge gradually as self-care lapses accumulate. This delayed pattern is medically plausible: patients with conditions like heart failure who miss medication doses or fail to monitor symptoms can experience rapid clinical deterioration within days. Clinical studies document that medication non-adherence substantially elevates mortality risk in the first month post-discharge (Simpson et al., 2006; Fitzgerald et al., 2011).

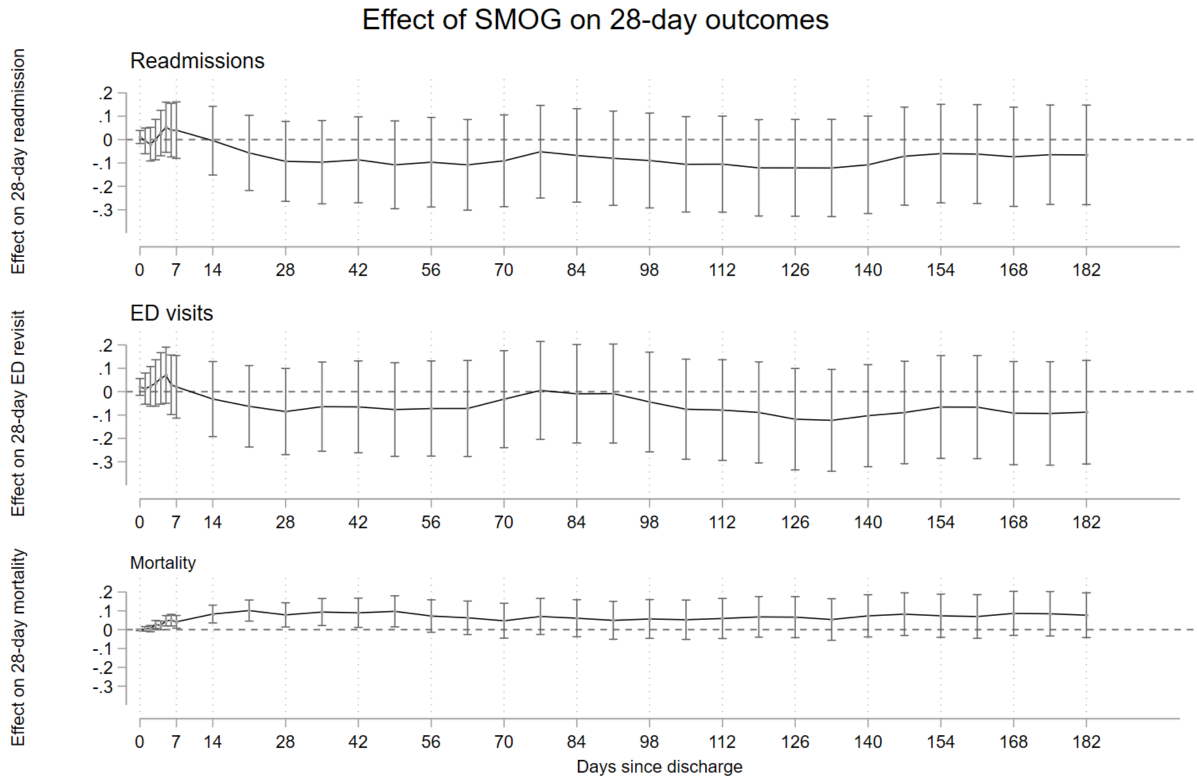
I re-estimate the main specification separately for mortality measured at different post-discharge horizons (0-7, 14, 21, 28, through 182 days). For each horizon t, the dependent variable is a binary indicator for cumulative mortality within t days of discharge. This non-parametric approach makes no assumptions about how the association evolves over time. While a Cox proportional hazards regression can be more statistically efficient, it imposes a proportional hazards assumption that the effect of complexity remains constant over time. As the results demonstrate, this assumption is violated. The series of separate estimates avoids this misspecification and provides a transparent view of the association's evolution. The analysis uses a subsample restricted to index admissions with no prior hospitalization in the preceding 182 days.

Figure 3 plots the SMOG coefficient over time for all three outcomes. The mortality pattern is consistent with an adherence mechanism rather than baseline frailty. The association is statistically insignificant for the first four days post-discharge (with the exception of day 3), emerges around day 5, peaks between weeks 3 and 7, and loses statistical significance by week 8. This inverted-U shape suggests risk concentrates in a vulnerable post-discharge window during which adherence lapses accumulate, rather than reflecting immediate baseline frailty. Coefficients for in-system readmissions and ED visits remain close to zero and statistically insignificant across all horizons. Selected horizons are reported in Appendix Table B6.

A third test examines whether complexity is a proxy for end-of-life status. Patients receiving palliative care may have complex advance directives documented in discharge notes, generating high mortality independent of communication quality. To test this, I exclude all patients with documented Do Not Resuscitate (DNR) or Do Not Intubate (DNI) orders. I identify these patients by searching the full text of all clinical notes from the hospital stay for explicit mentions of "DNR," "DNI," "Do Not Resuscitate," or "Do Not Intubate." Excluded patients have substantially higher baseline mortality (9.8 percent versus 1.2 percent in the retained sample). Appendix Table B3 shows that the association between SMOG and mortality remains positive and statistically significant in the restricted sample, though the magnitude is somewhat attenuated (0.08 pp, SE = 0.03). This provides evidence that end-of-life status, while potentially contributing to the baseline association, cannot fully explain the mortality gradient.

Treatment intensity provides another potential source of spurious correlation. Patients who undergo more procedures or receive more medications have longer, more technical discharge notes and independently face a higher mortality risk. I test whether this drives the association by augmenting the preferred specification with counts of lab tests, procedures, and prescriptions administered during hospitalization. If note complexity merely proxies for treatment intensity, the coefficient should attenuate substantially. Appendix Table B4 shows the opposite: the SMOG coefficient for mortality remains stable at 0.09 pp (SE = 0.03). The stability contrasts with the behavior of log

Figure 3: Temporal Pattern: Effect of SMOG on Post-Discharge Outcomes



Notes: Points represent coefficients and whiskers represent 95% confidence intervals from the preferred specification estimated separately at each time horizon. Sample restricted to index admissions (N = 135,675). Selected horizons reported in Appendix Table B6.

note length, whose coefficient for readmissions falls from 0.40 (SE = 0.18) to 0.23 and loses statistical significance after adding intensity controls. This pattern confirms that length partly proxies for severity in utilization outcomes while complexity does not.

A fifth concern is functional form misspecification. The preferred specification includes additive fixed effects for the patient's top three ICD diagnosis codes, which assumes separability: the effect of a secondary diagnosis of diabetes is constant across all primary diagnoses. This may be overly restrictive. A patient with acute myocardial infarction complicated by diabetes represents a clinically distinct entity from pneumonia complicated by diabetes. To test robustness, I replace the additive structure with a single concatenated fixed effect for each unique combination of the top three five-digit ICD codes. This non-parametric approach imposes no separability assumption and compares only patients with identical diagnosis profiles. The specification is exceptionally demanding: it drops all patients with unique diagnosis combinations (singletons), reducing the sample by 85 percent to 34,000 observations concentrated among complex cardiovascular and metabolic conditions. Despite this loss of power, the association remains positive and statistically significant (0.15 pp, SE = 0.08, p = 0.070). Appendix B.1 provides technical details and discusses the composition of the restricted sample.

Finally, I quantify the potential influence of remaining unobservables using the method of Oster (2019). This approach calculates the strength of selection on unobservables that would need to be relative to selection on observables to drive the estimated coefficient to zero. The method requires specifying R_{max}, the R-squared from a hypothetical regression that includes all possible controls. Following Oster's recommendation, I set $R_{max} = 1.3 \times R^2_{controlled}$, where $R^2_{controlled}$ is the R-squared from the preferred specification. The resulting δ is 1.50, indicating that unobservables would need to be 1.5 times as important as the full set of included controls to eliminate the mortality association. This exceeds the threshold of $\delta > 1$ commonly used to assess robustness, suggesting the association is unlikely to be driven entirely by omitted variables.

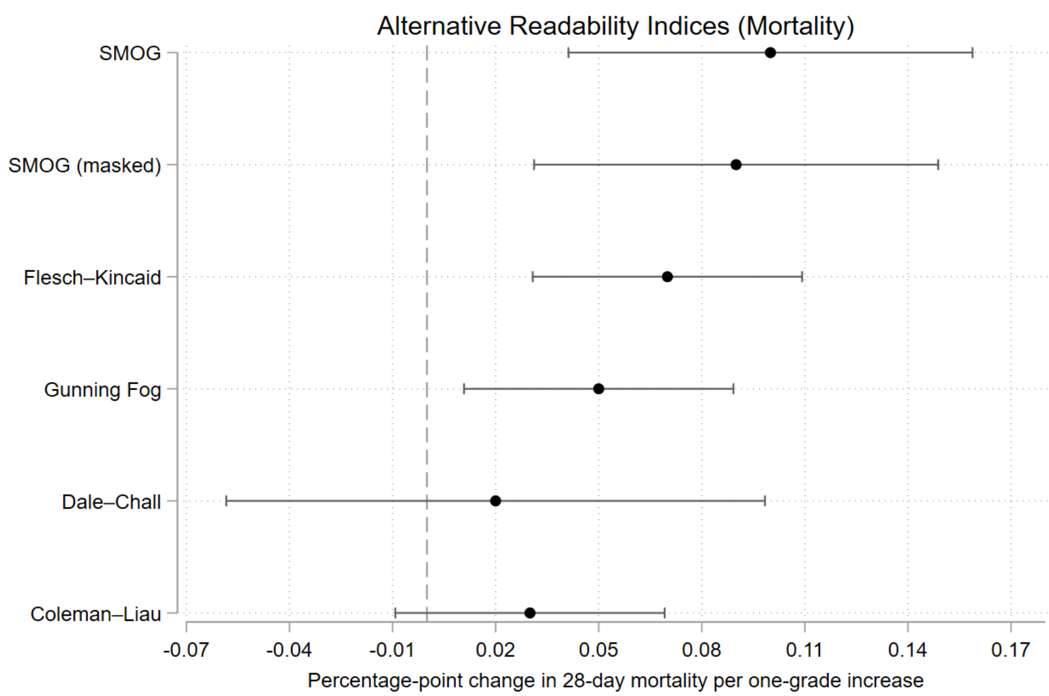
These six tests address the most plausible sources of confounding given the institutional context and data structure. Falsification tests and temporal analysis rule out that complexity proxies for baseline severity or immediate frailty. Excluding palliative patients demonstrates that end-of-life status contributes to but does not explain the association. Adding treatment intensity controls and allowing fully interacted diagnosis profiles verify that the association does not reflect specification choices or observable case complexity. The Oster bounds quantify what remains: unobservables would need to be 1.5 times as important as the extensive set of included controls to drive the coefficient to zero. Additional evidence that effects operate through discretionary writing choices comes from heterogeneity by template use: the mortality association is concentrated in non-template notes (0.12 pp, p = 0.001), with no detectable association in template-based notes (Appendix B.3). The combination of null falsification tests, delayed temporal onset, stable coefficients across demanding specifications, and favorable sensitivity bounds provides substantial support for interpreting the mortality association as reflecting communication barriers rather than unobserved confounding.

5.3 Robustness to Alternative Readability Measures

Two measurement concerns could threaten the interpretation of the baseline results. First, readability scores can be inflated by pharmacologic proper nouns (e.g., ”warfarin,” ”ibuprofen”) that add polysyllables without meaningfully increasing the cognitive burden of acting on an instruction. To probe this, I recompute a masked score that removes drug names before scoring. If the main result were an artifact of pharmacopoeia density, the mortality association should attenuate materially when drug names are masked.

Second, the result may be influenced by formula dependence. Readability indices emphasize different textual features, including syllables (SMOG), sentence length and syllables (Flesch-Kincaid), hard versus familiar words (Dale-Chall), and characters per word and sentence counts (Coleman-Liau). If the result reflects a genuine lexical-difficulty gradient rather than a SMOG-specific artifact, it should be visible, at least approximately, in alternative indices.

Figure 4: Alternative Readability Indices and 28-Day Mortality



Notes: Points are coefficients from the preferred specification; whiskers are 95% confidence intervals. Units are percentage-point changes in mortality per one-grade increase in the index. All specifications include the full control set and fixed effects. Standard errors clustered at the patient level. Full estimates in Appendix Tables B8 through B10.

Appendix Table B7 shows that masked SMOG preserves the pattern from Table 2: a positive, statistically significant association with 28-day mortality (0.09 pp per grade; SE = 0.03) and near-zero coefficients for readmission and ED revisits. Figure 4 visualizes the preferred-specification coefficients for six indices with 95 percent confidence intervals. The mortality estimates are uniformly positive and similar in magnitude across formulas: SMOG approximately 0.10 pp, Flesch-Kincaid approximately 0.07 pp, Gunning Fog approximately 0.05 pp. Dale-Chall and Coleman-Liau are smaller and less precisely estimated. Appendix Tables B8 through B10 report the underlying estimates and show that readmission and ED coefficients remain small and statistically indistinguishable from zero across formulas.

Some formulas yield smaller or less precise estimates because they measure different textual features. Dale-Chall hinges on membership in a familiar-word list; Coleman-Liau uses characters per word rather than syllables. These design choices alter sensitivity to clinical language (e.g., abbreviations, Latin terms, mixed narrative and imperative prose), thereby lowering the signal-to-noise ratio for the mechanism of interest: patient comprehension of actionable instructions. The persistence of the mortality gradient under masked SMOG and multiple alternative indices, alongside consistently null utilization effects, supports the interpretation that lexical difficulty, rather than drug-name density or a single scoring rule, tracks worse short-term survival.

5.4 Heterogeneity by Clinical Condition

The conceptual framework predicts that harm from linguistic complexity should concentrate among patients whose health is highly sensitive to self-care actions (high φ). This heterogeneity arises because for these patients, small errors in comprehension translate directly into adverse health outcomes through rapid physiological mechanisms. To test this prediction, I define a high-φ subsample consisting of Heart Failure and Chronic Obstructive Pulmonary Disease admissions. Clinical guidance emphasizes intensive post-discharge self-management for these conditions: daily weight monitoring and sodium restriction for Heart Failure; correct inhaler technique, maintenance therapy adherence, and action-plan execution for COPD. These demands are particularly acute during a well-documented high-risk window immediately after discharge (Heidenreich et al., 2022; Greene et al., 2015; Metra et al., 2023; Agustí et al., 2023). Consistent with the view that transitional care and patient understanding are margin-relevant, Heart Failure was included in the CMS Hospital Readmissions Reduction Program at its inception, and COPD was added in 2014.

As a contrast, I define a low-φ subsample of sepsis and cancer admissions. These cases exhibit comparable baseline 28-day mortality risk, but near-term outcomes are dominated by severity of organ dysfunction (sepsis) or disease biology and treatment toxicity (cancer), rather than incremental self-management (Courtright et al., 2020; Engoren et al., 2023; Ramchandran et al., 2013). The framework's prediction is sharp: if complexity impairs execution of self-care, effects should concentrate in high-φ conditions and attenuate for low-φ conditions.

I operationalize these subsamples using DRG and diagnosis codes: (i) Acute Heart Failure: DRGs 291–293 with ICD I50* (ICD-9: 428*); (ii) Acute COPD: DRGs 190–192 with ICD J44*, 492*, 496*; (iii) Sepsis: DRGs 870–872 with ICD A41* or septicemia codes; (iv) Cancer (metastatic): ICD C77–C79 (ICD-9: 196–199).[5]

Table 4 reports condition-specific estimates. Coefficients are in percentage points, consistent with the main results. Complexity effects concentrate in Heart Failure and do not generalize to sepsis or cancer, confirming the framework's prediction.

The estimates align with the prediction that harm concentrates where φ is high. In acute Heart Failure (Panel A), a one-grade-level increase in SMOG raises 28-day mortality by 0.95 pp (SE

[5] Complete coding logic in Appendix Table A2.

Table 4: Heterogeneity by Patient Cohort: Effects on 28-Day Outcomes

	Mortality	Readmission	ED Visits
Panel A: Acute Heart Failure			
SMOG	0.95***	0.66	1.07
	(0.29)	(0.66)	(0.69)
Log Note Length	-0.84	-1.26	-0.74
	(0.80)	(1.92)	(1.98)
Obs. / Clusters	4,219 / 3,018	4,219 / 3,018	4,219 / 3,018
Outcome mean (%)	3.10	22.96	25.83
Panel B: Acute COPD			
SMOG	0.96	1.14	1.65
	(0.60)	(2.13)	(2.35)
Log Note Length	-4.62**	-0.71	-1.77
	(2.22)	(7.01)	(7.08)
Obs. / Clusters	614 / 494	614 / 494	614 / 494
Outcome mean (%)	1.67	16.68	20.39
Panel C: Sepsis			
SMOG	-0.31	-0.16	-0.21
	(0.38)	(0.70)	(0.74)
Log Note Length	-0.52	4.33**	5.14***
	(1.14)	(1.86)	(1.95)
Obs. / Clusters	2,743 / 2,555	2,743 / 2,555	2,743 / 2,555
Outcome mean (%)	3.80	19.01	21.60
Panel D: Cancer (Metastatic)			
SMOG	0.30	-0.04	-0.27
	(0.36)	(0.59)	(0.60)
Log Note Length	-1.39	-0.24	-0.01
	(0.85)	(1.38)	(1.44)
Obs. / Clusters	7,498 / 5,417	7,498 / 5,417	7,498 / 5,417
Outcome mean (%)	7.70	27.82	31.19

Notes: Coefficients multiplied by 100 (percentage points). Standard errors clustered at the patient level in parentheses. All specifications include the full control set and fixed effects from the preferred specification.
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

= 0.29, p = 0.001). Scaling by the standard deviation of SMOG (1.4 grades), this translates to a 1.3 percentage point increase in mortality, nearly ten times the full-sample association of 0.14 percentage points. In acute COPD (Panel B), the point estimate is similar in magnitude at 0.96 pp, but imprecisely estimated (SE = 0.60, p = 0.111). The two high-φ conditions show qualitatively consistent magnitudes but differ in precision.

The COPD subsample is small (614 encounters; 494 clusters) and has low mortality, which limits statistical power. A standard power calculation yields a minimum detectable effect of 1.69 pp for 80% power at the 5% significance level. The point estimate of 0.96 pp falls below this threshold, consistent with underpowered testing rather than evidence of no effect. By contrast, the Heart Failure regression (MDE = 0.81 pp) detects effects at the observed magnitude.

The stronger Heart Failure result may also reflect differences in how information is executed. Core Heart Failure tasks—daily weight monitoring and sodium counting—are mediated primarily by written instructions, making them sensitive to linguistic complexity. In COPD, correct inhaler technique may depend more on in-person demonstration. A measure of written complexity should therefore have a greater impact on Heart Failure than on COPD patients, even as both conditions require intensive self-management.

In the low-φ conditions (sepsis and cancer, Panels C and D), SMOG coefficients are smaller, roughly one-third the magnitude of the high-φ estimates, and statistically insignificant. This pattern is consistent with near-term outcomes dominated by disease severity rather than self-management. Coefficients for readmissions and ED visits remain small and statistically insignificant across all four conditions, consistent with the measurement error in in-system utilization discussed in Section 5.1.

These results demonstrate that the mortality association concentrates where the conceptual framework predicts: among patients whose health production parameter φ is high, meaning their survival depends on executing complex self-care routines. The next section explores whether associations persist within narrower samples defined by specific instructional content.

5.5 Heterogeneity by Instructional Content

The baseline results and clinical condition heterogeneity establish that note complexity is associated with mortality. A remaining question is whether this reflects general lexical difficulty or whether specific types of instructions drive the association. To probe this, I employ structured large language model extraction to identify subsamples of notes containing specific self-care tasks. This approach serves two purposes. First, it demonstrates that the association persists within narrower action spaces: patients receiving the same type of instruction still face differential mortality based on how that instruction is written. Second, the extraction itself represents a methodological contribution, demonstrating how LLMs can systematically categorize clinical text at scale to enable content-specific analyses that are not feasible with traditional administrative data.

I apply a deterministic LLM extraction (Gemini via Vertex AI, temperature = 0) to flag whether four clinically salient instructions appear in each note: daily weight monitoring (present in 9.9 percent of notes), low-salt diet (1.8 percent), red-flag symptoms (44.5 percent), and stopping critical medications (19.8 percent). These tasks are selected because they represent non-trivial, high-stakes self-management decisions where comprehension failures could plausibly affect near-term outcomes: missed weight monitoring delays detection of fluid overload, dietary non-adherence accelerates decompensation, failure to recognize warning signs delays care-seeking, and medication errors can have immediate consequences. For each task, the model returns structured output indicating presence, supporting quotes, and a procedural complexity score (CICS). A manual review of a random sample of 100 extractions confirms 96% accuracy in identifying task presence.[6]

Importantly, this analysis does not identify the effect of instructional framing separate from content. Notes containing weight monitoring instructions are written for clinically distinct patients who require that monitoring. The value of this exercise lies in its interpretive nature: by restricting the analysis to notes that contain similar instructions, I narrow the action space and test whether lexical complexity (as measured by SMOG) predicts mortality even when task content is held approximately constant. If the baseline association were driven purely by heterogeneous task assignments rather than linguistic presentation, it should attenuate substantially within these content-defined subsamples.

Table 5 reports the SMOG coefficient from the preferred specification estimated separately within each task subsample. The mortality association persists in the two largest groups. Among notes with red-flag symptom instructions (N = 82,515), a one-grade increase in SMOG is associated with 0.08 pp higher mortality (SE = 0.04, p = 0.062). Among notes with weight monitoring

[6] Full prompt, schema, and validation details in Appendix A.3.

instructions (N = 18,661), the association is 0.18 pp (SE = 0.11, p = 0.089). These magnitudes are comparable to the full-sample estimate (0.10 pp) and smaller than the Heart Failure subsample (0.95 pp), consistent with task presence enriching for but not perfectly isolating adherence-sensitive patients. In the smaller low-salt and medication subsamples, estimates are imprecise. Coefficients for readmissions and ED visits remain insignificant across all task types, consistent with the baseline findings.

Table 5: Heterogeneity by Instructional Content

	(1) Weight Monitoring	(2) Low-Salt Diet	(3) Red-Flag Symptoms	(4) Stop Critical Medication
Panel A: 28-Day Mortality				
SMOG	0.18*	0.30	0.08*	-0.45
	(0.11)	(0.58)	(0.04)	(0.60)
Panel B: 28-Day Readmission				
SMOG	-0.25	-0.42	0.46	-0.07
	(0.29)	(1.30)	(0.58)	(1.47)
Panel C: 28-Day ED Visit				
SMOG	-0.13	-1.43	0.54	-0.64
	(0.28)	(1.31)	(0.55)	(1.35)
Observations	18,661	1,811	82,515	2,185

Notes: Each column reports the SMOG coefficient from the preferred specification estimated on the subsample of notes containing the specified instruction. Coefficients multiplied by 100 (percentage points). Standard errors clustered at patient level in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

I next test whether procedural complexity within instructions (CICS) moderates the SMOG gradient. Interacting SMOG with an indicator for high CICS (greater than or equal to 4) yields no consistent pattern: the weight monitoring interaction is 0.27 pp (SE = 0.17, p = 0.12), while interactions for other tasks are small and insignificant (Appendix Table B11). Combined with the weak correlation between SMOG and CICS (Appendix Table A5), this suggests the mortality gradient operates primarily through lexical difficulty rather than procedural branching logic.

Finally, for red-flag instructions, I test whether placement within the note (top, middle, or bottom third) affects the SMOG-mortality relationship. Neither the SMOG-by-position interactions nor position main effects are statistically significant (Appendix Table B12), providing no evidence that salience or document structure drives the observed gradient.

These results demonstrate that the association between note complexity and mortality persists even when restricting to patients who receive similar types of instructions. The persistence within content-defined subsamples supports the interpretation that how instructions are written, not merely what tasks are assigned, contributes to differential mortality risk.

5.6 Magnitude and External Benchmarks

A one-standard-deviation increase in instructional complexity is associated with a 0.14 percentage point increase in 28-day mortality, or 7.9 percent relative to the mean. For Heart Failure patients, the association is substantially larger: a one-standard-deviation increase in complexity is associated with a 1.18 percentage point increase in mortality, representing 38 percent relative to baseline. To assess whether these magnitudes are plausible, I compare them to estimates from two literatures: studies of hospital quality and healthcare access, as well as randomized evaluations of transition-of-care interventions.

Hospital quality and access studies document mortality effects of similar or larger magnitude. Chan et al. (2023) find that assignment to VA versus non-VA hospitals reduces 28-day mortality by 21.6 percent relative to baseline rates using OLS, and 46 percent using IV. Doyle et al. (2015) estimate that hospitals with 1 standard deviation higher spending reduce one-year mortality by 3.7 percentage points, approximately 10 percent of sample mortality. Doyle et al. (2019) find that a one standard deviation improvement in composite hospital quality measures reduces one-year mortality by approximately 7.5 percent. Hull (2020) examines the full distribution of hospital quality effects on 30-day mortality using the ambulance design and finds substantial heterogeneity across hospitals. Reassignment from a hospital at the 10th percentile to one at the 90th percentile of the quality distribution would reduce 30-day mortality by 3.6 to 4.5 percentage points, representing 22 to 27 percent of the sample mean. Card et al. (2009) show that Medicare eligibility reduces 28-day mortality by approximately 11 percent. My baseline estimate of 7.9 percent falls within this range, while the Heart Failure estimate of 38 percent is larger but involves a high-risk subsample during a documented vulnerable period.

Further relevant are randomized evaluations of interventions targeting discharge communication and transitional care. Project RED (Re-Engineered Discharge), a randomized trial that intensified discharge planning and patient education, decreased hospital utilization by about 30 percent within

30 days of discharge, though mortality effects were not statistically significant at conventional levels (Jack et al., 2009). The Care Transitions Intervention, which includes enhanced discharge preparation and post-discharge coaching, reduced rehospitalization rates by 30 percent at 30 days (Coleman et al., 2006). A systematic review and meta-analysis of medication adherence interventions in heart failure found that interventions improving adherence reduce mortality (relative risk 0.89) and hospital readmissions (odds ratio 0.79) (Ruppar et al., 2016). Clinical evidence indicates that medication-related adverse events account for 16 percent of early readmissions, with 40 percent deemed potentially preventable (Forster et al., 2003).

These benchmarks suggest that the magnitudes I estimate are consistent with a mechanism in which discharge communication quality affects patient outcomes. If intensive, resource-heavy interventions that redesign the entire discharge process can generate 20 to 30 percent reductions in mortality and larger reductions in utilization, finding that a one-standard-deviation increase in instructional complexity is associated with a 0.14 percentage point increase in 28-day mortality (7.9 percent relative to the mean), and substantially larger effects in the high-risk Heart Failure subsample, is well within the range of clinical and economic plausibility. Importantly, interventions to improve discharge instruction readability would be substantially less costly than policies that require increasing hospital spending by one standard deviation or redesigning entire care transition processes, while achieving mortality reductions of comparable magnitude.

6 Discussion

The pattern of results supports an interpretation that linguistic barriers in discharge instructions impede the execution of self-care routines. The mortality association emerges gradually over days rather than immediately, does not correlate with admission-time severity measures, persists after extensive controls for observable confounders, and concentrates where the conceptual framework predicts: among patients whose conditions require intensive daily self-management. The association remains when restricting to patients who receive similar instructional content, suggesting that linguistic presentation, not merely task assignment, contributes to differential outcomes. The specificity to mortality, combined with null findings for in-system utilization, likely reflects measurement error in readmissions and emergency department visits rather than the absence of effects, given that patients can seek care across multiple facilities in the study region.

These mortality reductions can be expressed in quality-adjusted life years to provide an illustrative welfare benchmark, though such calculations require strong assumptions. Given the study population's characteristics (mean age 62, mean Charlson Comorbidity Index 3.9), remaining life expectancy is substantially reduced relative to the general population. A conservative range of 5 to 10 quality-adjusted life years per death averted accounts for both remaining longevity and quality-of-life adjustments for this older, high-comorbidity cohort. Using a cost-effectiveness threshold of $100,000 per QALY (Vanness et al., 2021), a one-grade reduction in note complexity across the 189,859 admissions would avert approximately 190 deaths, corresponding to 950 to 1,900 QALYs, translating to $500 to $1,000 per patient. For Heart Failure admissions, a one-grade reduction across 4,219 discharges would avert approximately 40 deaths (200 to 400 QALYs), translating to $4,700 to $9,500 per Heart Failure patient. These calculations illustrate potential welfare magnitudes but rest on the assumption that the observed associations reflect causal effects of communication quality rather than residual confounding.

Several policy interventions follow. First, electronic health record systems could integrate real-time readability feedback for patient-facing sections, flagging complex language and suggesting simpler alternatives at the point of authoring. Given that discharge summaries are typically written by rotating physician trainees with limited training in patient communication, embedding such tools in the note editor would target the margin where variation arises. Second, hospitals could deploy automated plain-language translations of discharge instructions through patient portals, rendering identical clinical content at lower reading levels with visual scheduling aids. Recent evidence demonstrates that LLM-generated summaries can lower readability while preserving clinical accuracy (Zaretsky et al., 2024), supporting the technical feasibility of this approach. Third, medical education could integrate patient communication training into residency programs, with systematic feedback on note readability as part of competency assessment. Fourth, hospital quality monitoring systems could incorporate readability metrics alongside traditional measures such as infection rates and readmissions, creating institutional incentives for clear communication. Targeting these interventions to conditions with high self-care sensitivity (high φ) and task-rich notes would maximize returns.

The analysis has several important limitations that affect interpretation. The data come from a single large academic medical center, and whether these patterns generalize to community hospitals, rural settings, or different patient populations remains an open question. The identification strategy relies on extensive controls and fixed effects to address confounding; however, I am unable to link discharge notes to individual authors, which precludes constructing an author-level instrumental variable that would exploit the quasi-random assignment of patients to clinicians with different writing styles. The utilization outcomes—readmissions and emergency department visits—are measured only within the study hospital system, creating classical measurement error that attenuates any true associations toward zero; in contrast, mortality is comprehensively observed through Social Security death records. The primary complexity measure, SMOG, captures lexical difficulty through polysyllabic word density but does not directly measure procedural complexity or conditional logic within instructions, though supplemental analysis using instruction-level branching scores suggests that lexical difficulty is the dominant channel. Finally, while the falsification tests, temporal patterns, and heterogeneity analyses support a behavioral interpretation, the possibility of residual confounding from unobserved patient characteristics cannot be entirely ruled out. The results are best interpreted as documenting a robust and policy-relevant conditional association that is consistent with a causal mechanism, rather than as definitive evidence of causality. The convergence of evidence across multiple robustness checks, along with alignment with theoretical predictions and external benchmarks, strengthens confidence in a behavioral interpretation. However, formal causal claims would require experimental or quasi-experimental variation in note complexity.

7 Conclusion

The results identify communication clarity as a modifiable input in the health production function. While prior work has documented that patients struggle to comprehend discharge instructions, this paper provides the first large-scale evidence linking textual complexity to mortality. The findings complement experimental evidence from transition-of-care interventions, which show that improving discharge communication generates substantial reductions in adverse outcomes. Unlike those intensive, resource-heavy programs, improving note readability represents a scalable, low-cost margin for quality improvement.

The analysis leverages recent advances in natural language processing to measure complexity systematically and employs large language models to categorize instructional content at scale, demonstrating how computational linguistics can enable new empirical approaches in health economics. The concentration of effects among patients with high self-care sensitivity and the persistence within content-defined subsamples provide evidence that linguistic barriers, not merely clinical heterogeneity, contribute to differential survival.

These patterns suggest that targeted interventions to simplify discharge instructions, particularly for conditions requiring intensive self-care, could lead to meaningful reductions in post-discharge mortality at a modest cost. As healthcare systems increasingly seek high-return margins for quality improvement, communication represents an underexplored lever with potential for substantial welfare gains.

References

Abaluck, J. and J. Gruber (2011): "Choice Inconsistencies among the Elderly: Evidence from Plan Choice in the Medicare Part D Program," *American Economic Review*, 101, 1180–1210.

Agustí, A., B. R. Celli, G. J. Criner, D. Halpin, A. Anzueto, P. Barnes, J. Bourbeau, M. K. Han, F. J. Martinez, M. Montes de Oca, K. Mortimer, A. Papi, I. Pavord, N. Roche, S. Salvi, D. D. Sin, D. Singh, R. Stockley, M. V. López Varela, J. A. Wedzicha, and C. F. Vogelmeier (2023): "Global Initiative for Chronic Obstructive Lung Disease 2023 Report: GOLD Executive Summary," *American Journal of Respiratory and Critical Care Medicine*, 207, 819–837.

Arrow, K. J. (1963): "Uncertainty and the Welfare Economics of Medical Care," *American Economic Review*, 53, 941–73.

Axon, R. N., F. T. Penney, T. R. Kyle, J. Marsden, Y. Zhao, W. P. Moran, J. Zapka, and P. D. Mauldin (2014): "A Hospital Discharge Summary Quality Improvement Program Featuring Individual and Team-based Feedback and Academic Detailing," *The American Journal of the Medical Sciences*, 347, 472–477.

Becker, C., S. Zumbrunn, K. Beck, A. Vincent, N. Loretz, J. Müller, S. A. Amacher, R. Schaefert, and S. Hunziker (2021): "Interventions to Improve Communication at Hospital Discharge and Rates of Readmission: A Systematic Review and Meta-analysis," *JAMA Network Open*, 4, e2119346.

Bettinger, E. P., B. T. Long, P. Oreopoulos, and L. Sanbonmatsu (2012): "The Role of Application Assistance and Information in College Decisions: Results from the H&R Block Fafsa Experiment*," *The Quarterly Journal of Economics*, 127, 1205–1242.

Bhargava, S. and D. Manoli (2015): "Psychological Frictions and the Incomplete Take-Up of Social Benefits: Evidence from an IRS Field Experiment," *American Economic Review*, 105, 3489–3529.

Black, M. and C. M. Colford (2017): "Transitions of Care: Improving the Quality of Discharge Summaries Completed By Internal Medicine Residents," *MedEdPORTAL*, 10613.

Burns, S. T., N. Amobi, J. V. Chen, M. O'Brien, and L. A. Haber (2022): "Readability of Patient Discharge Instructions," *Journal of General Internal Medicine*, 37, 1797–1798.

Caplin, A. and M. Dean (2015): "Revealed Preference, Rational Inattention, and Costly Information Acquisition," *American Economic Review*, 105, 2183–2203.

Card, D., C. Dobkin, and N. Maestas (2009): "Does Medicare Save Lives?*," *Quarterly Journal of Economics*, 124, 597–636.

Chan, D. C., D. Card, and L. Taylor (2023): "Is There a VA Advantage? Evidence from Dually Eligible Veterans," *American Economic Review*, 113, 3003–3043.

Chandra, A. and D. Staiger (2007): "Productivity Spillovers in Health Care: Evidence from the Treatment of Heart Attacks," *Journal of Political Economy*, 115, 103–140.

Coleman, E. A., C. Parry, S. Chalmers, and S.-j. Min (2006): "The Care Transitions Intervention: Results of a Randomized Controlled Trial," *Archives of Internal Medicine*, 166, 1822.

Courtright, K. R., L. Jordan, C. M. Murtaugh, Y. Barrón, P. Deb, S. Moore, K. H. Bowles, and M. E. Mikkelsen (2020): "Risk Factors for Long-term Mortality and Patterns of End-of-Life Care Among Medicare Sepsis Survivors Discharged to Home Health Care," *JAMA Network Open*, 3, e200038.

Cutler, D. M. and M. McClellan (2001): "Productivity Change in Health Care," *American Economic Review*, 91, 281–286.

Doyle, J., J. Graves, and J. Gruber (2019): "Evaluating Measures of Hospital Quality: Evidence from Ambulance Referral Patterns," *The Review of Economics and Statistics*, 101, 841–852.

Doyle, J. J., J. A. Graves, J. Gruber, and S. A. Kleiner (2015): "Measuring Returns to Hospital Care: Evidence from Ambulance Referral Patterns," *Journal of Political Economy*, 123, 170–214.

Engoren, M., M. D. Maile, T. Seelhammer, R. E. Freundlich, and T. A. Schwann (2023): "Postdischarge Survival After Sepsis: A Cohort Study," *Anesthesia & Analgesia*, 137, 1216–1225.

Finkelstein, A. (2007): "The Aggregate Effects of Health Insurance: Evidence from the Introduction of Medicare," *The Quarterly Journal of Economics*, 122, 1–37.

Fitzgerald, A. A., J. D. Powers, P. M. Ho, T. M. Maddox, P. N. Peterson, L. A. Allen, F. A. Masoudi, D. J. Magid, and E. P. Havranek (2011): "Impact of Medication Nonadherence on Hospitalizations and Mortality in Heart Failure," *Journal of Cardiac Failure*, 17, 664–669.

Forster, A. J., H. J. Murff, J. F. Peterson, T. K. Gandhi, and D. W. Bates (2003): "The Incidence and Severity of Adverse Events Affecting Patients after Discharge from the Hospital," *Annals of Internal Medicine*, 138, 161–167.

Gabaix, X. (2014): "A Sparsity-Based Model of Bounded Rationality *," *The Quarterly Journal of Economics*, 129, 1661–1710.

Gentzkow, M., B. Kelly, and M. Taddy (2019): "Text as Data," *Journal of Economic Literature*, 57, 535–574.

Gentzkow, M. and J. Shapiro (2010): "What Drives Media Slant? Evidence From U.S. Daily Newspapers," *Econometrica*, 78, 35–71.

Glick, A. F., J. S. Farkas, A. L. Mendelsohn, A. H. Fierman, S. Tomopoulos, R. E. Rosenberg, B. P. Dreyer, J. Melgar, J. Varriano, and H. S. Yin (2019): "Discharge Instruction Comprehension and Adherence Errors: Interrelationship Between Plan Complexity and Parent Health Literacy," *The Journal of Pediatrics*, 214, 193–200.e3.

Goldberger, A. L., L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley (2000): "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals," *Circulation*, 101.

Greene, S. J., G. C. Fonarow, M. Vaduganathan, S. S. Khan, J. Butler, and M. Gheorghiade (2015): "The vulnerable phase after hospitalization for heart failure," *Nature Reviews Cardiology*, 12, 220–229.

Grossman, M. (1972): "On the Concept of Health Capital and the Demand for Health," *Journal of Political Economy*, 80, 223–255, publisher: University of Chicago Press.

Handel, B. R. and J. T. Kolstad (2015): "Health Insurance for "Humans": Information Frictions, Plan Choice, and Consumer Welfare," *American Economic Review*, 105, 2449–2500.

Heidenreich, P. A., B. Bozkurt, D. Aguilar, L. A. Allen, J. J. Byun, M. M. Colvin, A. Deswal, M. H. Drazner, S. M. Dunlay, L. R. Evers, J. C. Fang, S. E. Fedson, G. C. Fonarow, S. S. Hayek, A. F. Hernandez, P. Khazanie, M. M. Kittleson, C. S. Lee, M. S. Link, C. A. Milano, L. C. Nnacheta, A. T. Sandhu, L. W. Stevenson, O. Vardeny, A. R. Vest, and C. W. Yancy (2022): "2022 AHA/ACC/HFSA Guideline for the Management of Heart Failure," *Journal of the American College of Cardiology*, 79, e263–e421.

Heiss, F., A. Leive, D. McFadden, and J. Winter (2013): "Plan selection in Medicare Part D: Evidence from administrative data," *Journal of Health Economics*, 32, 1325–1344.

Hesselink, G., M. Zegers, M. Vernooij-Dassen, P. Barach, C. Kalkman, M. Flink, G. Ön, M. Olsson, S. Bergenbrant, C. Orrego, R. Suñol, G. Toccafondi, F. Venneri, E. Dudzik-Urbaniak, B. Kutryba, L. Schoonhoven, and H. Wollersheim (2014): "Improving patient discharge and reducing hospital readmissions by using Intervention Mapping," *BMC Health Services Research*, 14, 389.

Ho, K., J. Hogan, and F. Scott Morton (2017): "The impact of consumer inattention on insurer pricing in the Medicare Part D program," *The RAND Journal of Economics*, 48, 877–905.

Horwitz, L. I., G. Y. Jenq, U. C. Brewster, C. Chen, S. Kanade, P. H. Van Ness, K. L. B. Araujo, B. Ziaeian, J. P. Moriarty, R. L. Fogerty, and H. M. Krumholz (2013): "Comprehensive quality of discharge summaries at an academic medical center," *Journal of Hospital Medicine*, 8, 436–443.

Hull, P. (2020): "Estimating Hospital Quality with Quasi-Experimental Data," *Unpublished*.

Jack, B. W., V. K. Chetty, D. Anthony, J. L. Greenwald, G. M. Sanchez, A. E. Johnson, S. R. Forsythe, J. K. O'Donnell, M. K. Paasche-Orlow, C. Manasseh, S. Martin, and L. Culpepper (2009): "A Reengineered Hospital Discharge Program to Decrease Rehospitalization: A Randomized Trial," *Annals of Internal Medicine*, 150, 178–187.

Johnson, A., L. Bulgarelli, T. Pollard, B. Gow, B. Moody, S. Horng, L. A. Celi, and R. Mark (2024): "MIMIC-IV," .

Johnson, A., T. Pollard, S. Horng, L. A. Celi, and R. Mark (2023a): "MIMIC-IV-Note: Deidentified free-text clinical notes," .

Johnson, A. E. W., L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, L.-W. H. Lehman, L. A. Celi, and R. G. Mark (2023b): "MIMIC-IV, a freely accessible electronic health record dataset," *Scientific Data*, 10, 1.

Kripalani, S., C. N. Theobald, B. Anctil, and E. E. Vasilevskis (2014): "Reducing Hospital Readmission Rates: Current Strategies and Future Directions," *Annual Review of Medicine*, 65, 471–485.

Legault, K., J. Ostro, Z. Khalid, P. Wasi, and J. J. You (2012): "Quality of discharge summaries prepared by first year internal medicine residents," *BMC Medical Education*, 12, 77.

Leonard Grabeel, K., J. Russomanno, S. Oelschlegel, E. Tester, and R. E. Heidel (2018): "Computerized versus hand-scored health literacy tools: a comparison of Simple Measure of Gobbledygook (SMOG) and Flesch-Kincaid in printed patient education materials," *Journal of the Medical Library Association*, 106.

Madrian, B. C. and D. F. Shea (2001): "The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior," *The Quarterly Journal of Economics*, 116, 1149–1187.

Mc Laughlin, H. (1969): "SMOG Grading-a New Readability Formula," *Journal of Reading*, 12, 639–646.

Metra, M., M. Adamo, D. Tomasoni, A. Mebazaa, A. Bayes-Genis, M. Abdelhamid, S. Adamopoulos, S. D. Anker, J. Bauersachs, Y. Belenkov, M. Böhm, T. B. Gal, J. Butler, A. Cohen-Solal, G. Filippatos, F. Gustafsson, L. Hill, T. Jaarsma, E. A. Jankowska, M. Lainscak, Y. Lopatin, L. H. Lund, T. McDonagh, D. Milicic, B. Moura, W. Mullens, M. Piepoli, M. Polovina, P. Ponikowski, A. Rakisheva, A. Ristic, G. Savarese, P. Seferovic, R. Sharma, T. Thum, C. G. Tocchetti, S. Van Linthout, C. Vitale, S. Von Haehling, M. Volterrani, A. J. Coats, O. Chioncel, and G. Rosano (2023): "Pre-discharge and early post-discharge management of patients hospitalized for acute heart failure: A scientific statement by the Heart Failure Association of the ESC," *European Journal of Heart Failure*, 25, 1115–1131.

Ming, D., K. Zietlow, Y. Song, H. Lee, and A. Clay (2019): "Discharge summary training curriculum: a novel approach to training medical students how to write effective discharge summaries," *The Clinical Teacher*, 16, 507–512.

Oprea, R. (2020): "What Makes a Rule Complex?" *American Economic Review*, 110, 3913–3951.

Oster, E. (2019): "Unobservable Selection and Coefficient Stability: Theory and Evidence," *Journal of Business & Economic Statistics*, 37, 187–204.

Ramchandran, K. J., J. W. Shega, J. Von Roenn, M. Schumacher, E. Szmuilowicz, A. Rademaker, B. B. Weitner, P. D. Loftus, I. M. Chu, and S. Weitzman (2013): "A predictive model to identify hospitalized cancer patients at risk for 30-day mortality based on admission criteria via the electronic medical record," *Cancer*, 119, 2074–2080.

Ruppar, T. M., P. S. Cooper, D. R. Mehr, J. M. Delgado, and J. M. Dunbar-Jacob (2016): "Medication Adherence Interventions Improve Heart Failure Mortality and Readmission Rates: Systematic Review and Meta-Analysis of Controlled Trials," *Journal of the American Heart Association*, 5, e002606.

Russell, S., N. Jacobson, and A. Pavlic (2024): "Improving Patient Understanding of Emergency Department Discharge Instructions," *Western Journal of Emergency Medicine*, 25.

Simpson, S. H., D. T. Eurich, S. R. Majumdar, R. S. Padwal, R. T. Tsuyuki, J. Varney, and J. A. Johnson (2006): "A meta-analysis of the association between adherence to drug therapy and mortality," *BMJ*, 333, 15.

Vanness, D. J., J. Lomas, and H. Ahn (2021): "A Health Opportunity Cost Threshold for Cost-Effectiveness Analysis in the United States," *Annals of Internal Medicine*, 174, 25–32.

Wallace, L. S. and E. S. Lennon (2004): "American Academy of Family Physicians patient education materials: can patients read them?" *Family Medicine*, 36, 571–574.

Wang, L.-W., M. J. Miller, M. R. Schmitt, and F. K. Wen (2013): "Assessing readability formula differences with written health information materials: Application, results, and recommendations," *Research in Social and Administrative Pharmacy*, 9, 503–516.

You, P., J. Liu, L. Moist, K. Fung, and J. E. Strychowsky (2022): "Improving Timeliness in Surgical Discharge Summary Distribution: A Quality Improvement Initiative," *OTO Open*, 6, 2473974X221134106.

Zaretsky, J., J. M. Kim, S. Baskharoun, Y. Zhao, J. Austrian, Y. Aphinyanaphongs, R. Gupta, S. B. Blecker, and J. Feldman (2024): "Generative Artificial Intelligence to Transform Inpatient Discharge Summaries to Patient-Friendly Language and Format," *JAMA network open*, 7, e240357.

A Sample Construction and Text Processing

This appendix details the construction of the analysis sample, the processing of raw text data, and the creation of all variables used in the paper.

A.1 Sample Construction

The sample construction process, which details the sequential application of exclusion criteria to arrive at the final analysis sample, is outlined in Table A1. The initial cohort is derived from the MIMIC-IV and MIMIC-IV-Note databases. An "index admission" is defined as a hospital stay that was not preceded by another admission for the same patient within the prior 28 days.

Table A1: Sample Construction

Restriction	Admissions Dropped	Admissions Remaining
Initial cohort of index admissions with non-null discharge note	—	262,755
Exclude patient deaths during the index admission	92	262,663
Exclude notes with fewer than 100 characters	6,938	255,725
Exclude notes where SMOG score could not be calculated	6,758	248,967
Exclude notes with outlier sentence length (bottom 1% and top 1%)	4,955	244,012
Exclude discharges to hospice (location or mention of "hospice" in text)	2,705	241,307
Exclude discharges against medical advice	1,313	239,994
Exclude discharge notes in Non-English language	116	239,878
Final Analysis Sample	—	**239,878**

Notes: The table details the sequential application of exclusion criteria to arrive at the final analysis sample. The initial cohort is derived from the MIMIC-IV and MIMIC-IV-Note databases. An "index admission" is defined as a hospital stay that was not preceded by another admission for the same patient within the prior 28 days.

Table A2: Cohort Coding Rules

Cohort	Coding rule (any of top six diagnosis fields unless otherwise noted)
Acute CHF	DRGs 291–293 AND ICD I50* (ICD-9: 428*).
Acute COPD	DRGs 190–192 AND ICD J44*, 492*, 496*, or 49121.
Sepsis	DRGs 870–872 AND (A41* OR any of {99591, 99592, 78552, A419, R6520, R6521}).
Cancer (metastatic)	ICD C77–C79 (ICD-9: 196–199).

Notes: This table defines the coding rules used to identify clinical condition subsamples for heterogeneity analysis. DRG = Diagnosis-Related Group; ICD = International Classification of Diseases.

A.2 Text Processing and Readability Measurement

This section details the multi-step pipeline used to process raw clinical text and generate the variables for the analysis. The full code is available in the replication package.

Section Extraction. The process begins by extracting the patient-facing instruction text from the full discharge summary note. I use a regular expression to identify canonical headers such as "Discharge Instructions" and "Follow-up Instructions." If multiple matches are found, the final occurrence is selected to prioritize the conclusive guidance given to the patient. The extracted segment begins after the header and ends at the onset of subsequent standard headers to minimize contamination from physician-facing text.

Text Cleaning and Normalization. The extracted raw text then undergoes a rigorous cleaning process. This involves several steps: (1) fixing encoding errors and normalizing Unicode punctuation to ASCII; (2) removing URLs, phone numbers, and placeholder characters (e.g., ‘_____‘); (3) stripping boilerplate banner lines (e.g., "Division of Vascular and Endovascular Surgery"); and (4) harmonizing various bullet point formats (e.g., ‘*‘, ‘ ‘, ‘1)‘) into a standard-dashed list format.

Sentence Preparation and Segmentation. To ensure an accurate sentence count, which is a critical input for the SMOG index, the cleaned text is prepared for segmentation. Bulleted and numbered list items are converted into complete prose sentences. A terminal period is added conservatively to list items that end without punctuation. Light punctuation repair is also performed, for example, by replacing semicolons that introduce new clauses with periods. The prepared text is then segmented into sentences using a rule-based sentencizer from the ‘spaCy‘ library.

Readability Index Calculation. The primary explanatory variable, complexity, is measured using the SMOG index, calculated on the fully cleaned and segmented text. The SMOG formula is:

$$\text{SMOG} \;=\; 1.0430 \times \sqrt{\frac{30 \times \text{Number of Polysyllabic Words}}{\text{Number of Sentences}}} \;+\; 3.1291$$

As robustness checks, I also compute several alternative indices, including Flesch–Kincaid Grade Level, Flesch Reading Ease, and Gunning Fog.

Drug Name Identification and Masking. A potential concern is that note complexity could be mechanically correlated with patient severity if sicker patients receive more medications, and drug names are often long, polysyllabic words. To address this, I create a "drug-masked" version of the text. I use a natural language processing pipeline built with ‘spaCy‘ and ‘scispaCy‘ that employs a pre-trained Named Entity Recognition (NER) model to identify drug and chemical entities. These entities are then linked to the RxNorm vocabulary to improve accuracy. Identified drug names are replaced with a generic ‘<DRUG>‘ token. This process creates a ‘masked_text‘ variable, from which I calculate a ‘smog_masked‘ score to test if the paper's main results are robust to excluding complexity driven by medication names.

Template Detection. To identify notes that are likely boilerplate templates, I use a machine learning approach. First, I canonicalize the text by converting it to lowercase and removing numbers and punctuation. I then use a ‘SentenceTransformer‘ model to embed each canonicalized note into a high-dimensional vector. Finally, I use the ‘DBSCAN‘ clustering algorithm to identify groups of notes that are highly similar in this vector space, flagging them as templates. This allows for heterogeneity analysis based on whether the note was likely free-form or template-based.

A.3 LLM-Based Task Extraction

Model and settings. I use Gemini 2.5 Flash via Google's Vertex AI API with deterministic decoding (temperature = 0), JSON-only output format, and safety filters disabled to process clinical content without blocking.
Batch processing. The extraction was performed via Google's Vertex AI batch processing API to handle the full sample efficiently. All processing used deterministic settings to ensure reproducibility.
Prompt design. The model receives a system instruction defining four clinical tasks and returns structured output indicating: (i) whether each task is present in the note, (ii) the verbatim quote supporting that determination, (iii) a procedural complexity score (CICS, scale 1–5) measuring the number of conditionals and steps, and (iv) for red-flag symptoms, the location within the document. This structured format enables systematic analysis of both task content and linguistic presentation across 239,878 discharge notes. The full system instruction follows:

> You are a specialized AI, an expert in clinical document analysis. Your mission is to meticulously analyze patient discharge instructions and extract key tasks into a structured JSON object. You must adhere to the rules and schema provided with absolute precision.
>
> You MUST respond with only a valid JSON object and nothing else. Do not include any explanatory text, markdown formatting, or apologies.
>
> **Rules of Analysis**
>
> For each discharge note, you will analyze the presence, evidence, and complexity of four key patient tasks.
>
> **1. Task Definitions:**
>
> - **Weight Monitoring:** Look for any instruction for the patient to weigh themselves, monitor their weight daily, or record their weight.
> - **Low Salt Diet:** Look for any instruction to follow a low salt, low sodium, salt-restricted, or specific gram-limited (e.g., "2 gm") sodium diet.
> - **Red Flag Symptoms:** Look for instructions telling the patient to call a doctor or go to the ER for a list of serious symptoms (e.g., chest pain, shortness of breath, weakness, numbness, new bleeding, fever).
>   - For the location key, determine where the primary instruction appears: "top" (first 25%), "middle" (25%–75%), or "bottom" (final 25%).
> - **Stop Critical Medication:** Look for an explicit instruction to stop, hold, or discontinue a critical medication. Critical medications include (but are not limited to): blood thinners (Coumadin, Aspirin), heart/blood pressure medications (Metoprolol, Lisinopril), diuretics (Lasix), insulin, and steroids. Exclude routine, non-critical instructions like stopping Tylenol or temporary pain medication.
>
> **2. Complexity Assessment:**
>
> For each task that is present, assess the complexity of the instruction using the Clinical Instruction Complexity Scale (CICS):
>
> - **CICS 1 (Very Simple):** A short, direct, unconditional command. *Example:* "Weigh yourself daily."

- **CICS 2 (Simple with Quantifier):** A direct command that includes a specific number, measurement, or simple list. *Example:* "Adhere to 2 gm sodium diet."
- **CICS 3 (Moderate):** A longer sentence or a command that includes a list of 3+ distinct items. *Example:* "Return to the ER for chest pain, palpitations, weakness, or numbness."
- **CICS 4 (Complex):** A command containing a conditional element (if/then) OR a brief rationale. *Example:* "Do NOT take any more NSAIDs as it may affect your kidney function."
- **CICS 5 (Very Complex):** A multi-part conditional instruction or a command embedded in a long paragraph requiring significant inference. *Example:* "Please restart in 2 days if diarrhea resolving. Otherwise discuss with your PCP."

Required JSON schema. The model returns a structured object with the hospital admission ID and a nested task analysis section. Each task includes a binary presence indicator, the supporting text quote, and the complexity score. For red-flag symptoms, an additional location field indicates document placement. Tasks not present in the note have null values for quote and complexity fields:

```
{
  "hadm_id": 12345,
  "task_analysis": {
    "weight_monitoring": {
      "present": 1,
      "reasoning_quote": "Weigh yourself every morning...",
      "complexity_cics": 4
    },
    "low_salt_diet": {
      "present": 0,
      "reasoning_quote": null,
      "complexity_cics": null
    },
    "red_flag_symptoms": {
      "present": 1,
      "reasoning_quote": "Call for chest pain...",
      "location": "bottom",
      "complexity_cics": 3
    },
    "stop_critical_medication": {
      "present": 0,
      "reasoning_quote": null,
      "complexity_cics": null
    }
  }
}
```

Generation configuration:

```
max_output_tokens: 8192
temperature: 0
response_mime_type: application/json
safety_settings: BLOCK_NONE for all categories
```

Validation. I manually reviewed a random sample of 100 extractions to assess accuracy. The model correctly identified task presence with 96% accuracy.

A.4 Descriptive Statistics for LLM-Extracted Variables

Table A3: Prevalence of LLM-Identified Tasks in Discharge Notes

Task indicator	Freq.	Percent (%)	Cum. (%)
Weight present	23,560	9.88	100.00
Not present	214,797	90.12	90.12
Low-salt present	4,373	1.83	100.00
Not present	233,984	98.17	98.17
Stop/hold medication present	47,251	19.82	100.00
Not present	191,106	80.18	80.18
Red-flag present	106,049	44.49	100.00
Not present	132,308	55.51	55.51
Total notes	238,357	100.00	

Notes: Percentages computed on the full corpus (N = 238,357).

Table A4: CICS Distribution (Within-Instruction Procedural Complexity)

	Share within task (%)				
Task (subsample size)	CICS=1	CICS=2	CICS=3	CICS=4	CICS=5
Weight (N=23,560)	1.09	9.15	15.93	70.12	3.72
Low-salt (N=4,373)	26.25	61.08	4.23	8.35	0.09
Red-flag (N=106,049)	0.33	1.42	33.00	52.74	12.51
Stop/hold med (N=47,251)	20.07	10.47	10.00	55.52	3.94

Notes: Shares computed within each task-present subsample. Row headers report the corresponding sample size N.

Table A5: Mean SMOG by CICS Tier and Task

	CICS=1	CICS=2	CICS=3	CICS=4	CICS=5
Weight (mean SMOG)	10.52	10.12	10.30	10.50	10.66
Low-salt (mean SMOG)	10.48	10.66	10.79	10.69	11.58
Red-flag (mean SMOG)	11.28	10.40	10.94	11.43	11.02
Stop/hold med (mean SMOG)	10.85	10.85	10.69	11.14	11.37

Notes: Entries are sample means of SMOG within each CICS tier (task-present subsamples). Underlying sample sizes by tier appear in Table A4.

B Additional Results and Robustness Checks

Table B1: Number of Levels for High-Dimensional Fixed Effects

Fixed Effect Variable	Number of Levels
Clinical Fixed Effects	
Top 5-Digit ICD Code 1	5,512
Top 5-Digit ICD Code 2	5,406
Top 5-Digit ICD Code 3	5,860
DRG (HCFA)	737
Last Clinical Service	18
Institutional & Logistical Fixed Effects	
Admitting Provider	1,313
Discharge Hour	24
Anchor Year Group	4
Admission Location	11
Discharge Location	10
Admission Type	9
Patient Demographic Fixed Effects	
Language	25
Race	33
Insurance	5

Notes: This table reports the number of unique levels (groups) for each of the high-dimensional fixed effects absorbed in the preferred specification (Column 4 of Table 2). These fixed effects are used to control for unobserved heterogeneity at the clinical, institutional, and patient levels.

B.1 Robustness to a Non-Parametric Specification of Clinical Severity

This section tests robustness to a fully non-parametric specification of clinical severity. The main analysis includes additive fixed effects for the patient's top three five-digit ICD codes, which assumes that the effect of a secondary diagnosis is constant across all primary diagnoses. I replace this with a single concatenated fixed effect for each unique combination of the top three codes, making no assumptions about separability. This ensures that identifying variation comes from comparing patients with identical diagnosis profiles.

This is an exceptionally demanding specification. A primary consequence of its implementation is a significant reduction in the estimation sample, as any patient whose specific combination of top-three ICD codes is unique in the dataset (a singleton) is dropped from the analysis. As shown in Table B2, this reduces the effective sample size by over 85%, from approximately 240,000 to 34,000 observations. The analysis is therefore concentrated on a non-random subsample of patients with the most frequently observed combinations of comorbidities. This subsample is composed largely of patients with complex cardiovascular and metabolic conditions, such as combinations of coronary artery disease (ICD-9 414.01), various forms of heart failure (ICD-9 428.23, ICD-10 I50.33), hypertension (ICD-9 401.9), and acute kidney injury (ICD-9 584.9, ICD-10 N17.9).

Table B2 presents the results of this robustness check. For each of the three main outcomes, Column (1) reports the estimates from the demanding concatenated fixed-effects model, while Column (2) reports the estimates from the standard additive fixed-effects model estimated on the identical, restricted sample. This pairing allows for a direct assessment of the sensitivity of the estimates to the choice of severity control, holding the sample constant.

The central finding of the paper proves robust. Panel A shows that the association between note complexity and 28-day mortality is stable across specifications. In the concatenated model (Column 1), the point estimate is 0.15 (p=0.070). In the additive model on the same restricted sample (Column 2), the point estimate is 0.13 (p=0.056). The stability of the coefficient's magnitude and statistical significance across these two specifications, despite the dramatic loss of precision in the small subsample, provides strong evidence for the robustness of the paper's headline result.

The results for the other outcomes are sensitive to the choice of specification within this subsample. For 28-day readmissions (Panel B), the concatenated specification in Column (1) yields a positive and statistically significant coefficient of 0.46 (p=0.048). However, this association is not robust. Column (2) shows that when the standard additive fixed-effects model is estimated on the identical sample, the point estimate shrinks by over 30% to 0.31 and is only significant at the 10% level (p=0.094). A similar pattern holds for 28-day ED visits (Panel C), where neither specification yields a statistically significant association.

Taken together, this analysis yields two key insights. First, the paper's central finding - the positive association between note complexity and mortality - is stable even under an exceptionally stringent, non-parametric specification for clinical severity. Second, the null associations for readmissions and ED visits from the main, full-sample analysis appear to be the most credible estimates, as the alternative specifications produce fragile and specification-sensitive results within a highly selected subsample.

Table B2: Robustness to a Non-Parametric Specification of Clinical Severity

	(1)	(2)
ICD Specification:	Concatenated	Additive
Sample:	Restricted	Restricted
Panel A: 28-Day Mortality		
SMOG Index (pp change)	0.15 *	0.13 *
	(0.08)	(0.07)
Panel B: 28-Day Readmission		
SMOG Index (pp change)	0.46 **	0.31 *
	(0.23)	(0.18)
Panel C: 28-Day ED Visit		
SMOG Index (pp change)	0.27	0.10
	(0.22)	(0.18)
Observations	33,822	33,822

Notes: This table assesses the robustness of the main findings to a highly demanding fixed-effect structure on a restricted sample. Column (1) replaces the separate, additive fixed effects for the top three ICD codes with a single, concatenated fixed effect for each unique combination of the top three 5-digit ICD-9/10 codes. Column (2) uses the standard, additive fixed effects but is estimated on the exact same, restricted sample as Column (1). All specifications include the full set of controls from the main analysis, including admission number, but exclude admitting provider fixed effects. Coefficients have been multiplied by 100 for interpretation as percentage point changes. Standard errors, reported in parentheses, are clustered at the patient level. *** p<0.01, ** p<0.05, * p<0.10.

B.2 Further Robustness Checks

Table B3: Main Specification on Sample Excluding Patients with DNR/DNI Orders

	(1) 28-Day Mortality	(2) 28-Day Readmission	(3) 28-Day ED Visit
SMOG	0.064**	-0.042	-0.073
	(0.024)	(0.077)	(0.076)
Log Note Length	-0.201***	0.347*	0.422**
	(0.056)	(0.182)	(0.182)
Observations	180,559	180,559	180,559
Full Controls	✓	✓	✓

Notes: This table presents estimates from the preferred specification (Column 4 of Table 2) on the subsample of patients without a documented Do-Not-Resuscitate (DNR) or Do-Not-Intubate (DNI) order. All coefficients are multiplied by 100. Standard errors, in parentheses, are clustered at the patient level. *** p<0.01, ** p<0.05, * p<0.10

Table B4: Robustness to Controlling for In-Hospital Treatment Intensity

	(1) 28-Day Mortality	(2) 28-Day Readmission	(3) 28-Day ED Visit
SMOG	0.099***	-0.035	-0.045
	(0.027)	(0.075)	(0.074)
Log Note Length	-0.353***	0.232	0.301*
	(0.064)	(0.179)	(0.179)
Observations	189,859	189,859	189,859
Full Controls	✓	✓	✓
Treatment Intensity Controls	✓	✓	✓

Notes: This table presents estimates from the preferred specification (Column 4 of Table 2) augmented with a rich set of controls for in-hospital treatment intensity. These controls include the number of lab tests (total and STAT), number of procedures, and number of prescriptions. All coefficients are multiplied by 100. Standard errors, in parentheses, are clustered at the patient level.
*** p<0.01, ** p<0.05, * p<0.10

B.3 Heterogeneity by Template Use

To examine whether the association operates through discretionary writing choices or standardized text, I classify notes as template-based using a machine learning approach detailed in Appendix A.2. I canonicalize each note, embed the text into a high-dimensional vector space using a sentence transformer model, and apply DBSCAN clustering to identify groups of highly similar notes. Notes assigned to clusters are flagged as template-based; singletons are classified as non-template.

This classification yields 86,209 template-based notes (35.9 percent) and 153,669 non-template notes (64.1 percent). Template-based notes exhibit substantially lower baseline mortality (0.68 percent versus 2.00 percent), indicating strong selection: templates are systematically used for lower-risk cases. Mean complexity differs slightly (SMOG 10.99 versus 11.21), though distributions substantially overlap (standard deviations 1.37 and 1.42).

Table B5 presents the preferred specification estimated separately by template status. The association between SMOG and mortality is concentrated in non-template notes. For non-template notes (Panel A), a one-grade increase in SMOG is associated with 0.12 percentage points higher mortality (SE = 0.04, p = 0.001). For template-based notes (Panel B), the association is statistically indistinguishable from zero (0.04 pp, SE = 0.04, p = 0.29).

Table B5: Heterogeneity by Template Use

	(1) 28-Day Mortality	(2) 28-Day Readmission	(3) 28-Day ED Visit
Panel A: Non-Template Notes			
SMOG	0.119***	-0.042	-0.010
	(0.037)	(0.093)	(0.092)
Log Note Length	-0.449***	0.375	0.435*
	(0.100)	(0.246)	(0.245)
Outcome mean (%)	2.00	18.55	17.55
Observations	118,121	118,121	118,121
Clusters	71,043	71,043	71,043
Panel B: Template-Based Notes			
SMOG	0.042	-0.157	-0.133
	(0.040)	(0.163)	(0.165)
Log Note Length	-0.216***	0.333	0.361
	(0.083)	(0.340)	(0.347)
Outcome mean (%)	0.68	14.39	14.77
Observations	65,089	65,089	65,089
Clusters	51,193	51,193	51,193

Notes: Each panel reports estimates from the preferred specification (Column 4 of Table 2) estimated on the indicated subsample. Template-based notes are identified using DBSCAN clustering on sentence-transformer embeddings as described in Appendix A.2. Coefficients multiplied by 100 (percentage points). Standard errors clustered at patient level in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

This pattern may reflect that discretionary writing exhibits more variation after conditioning on observables: residual SMOG variation after absorbing fixed effects is 40 percent larger in non-template notes (SD 1.21 versus 0.85). Alternatively, template patients may be less sensitive to communication quality due to lower baseline risk. The concentration of effects in non-template notes is consistent with linguistic barriers mattering most where communication is least standardized.

Table B6: Effect of SMOG on Outcomes at Selected Post-Discharge Horizons

Horizon (days)	Readmission	ED revisit	Mortality
0	0.011	0.020	−0.001
	(0.014)	(0.018)	(0.002)
1	−0.006	0.013	0.004
	(0.028)	(0.034)	(0.006)
2	−0.020	0.023	0.006
	(0.037)	(0.043)	(0.008)
7	0.040	0.021	0.042**
	(0.062)	(0.068)	(0.017)
14	−0.004	−0.031	0.083***
	(0.075)	(0.082)	(0.024)
28	−0.093	−0.085	0.078**
	(0.087)	(0.094)	(0.033)
56	−0.097	−0.072	0.072*
	(0.098)	(0.104)	(0.044)
84	−0.068	−0.009	0.061
	(0.102)	(0.107)	(0.050)
182	−0.065	−0.088	0.077
	(0.109)	(0.113)	(0.061)

Notes: Entries are coefficients with standard errors in parentheses. Stars denote ***p<0.01, **p<0.05, *p<0.10 (normal approximation). $N = 135{,}675$.

B.4 Robustness to Alternative Readability Measures

Table B7: Preferred Specification Using Masked Readability

	Mortality	Readmission	ED Visit
SMOG (masked)	0.091***	-0.055	-0.038
	(0.028)	(0.076)	(0.075)
Log Note Length	-0.337***	0.397**	0.432**
	(0.067)	(0.183)	(0.183)
Observations	189,859	189,859	189,859
Clusters (patients)	109,112	109,112	109,112

Notes: Linear probability models with the preferred control set and the full set of high-dimensional fixed effects as in Table 2. Entries are coefficients multiplied by 100 (percentage points). Robust standard errors (in parentheses) are clustered at the patient level. *Outcomes*: 28-day all-cause mortality (col. 1), in-system readmission (col. 2), and in-system ED visit (col. 3).
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table B8: Alternative Readability Indices: 28-Day Mortality

	SMOG	Masked	Dale–Chall	Flesch–Kincaid	Gunning Fog	Coleman–Liau
Coefficient	0.10***	0.09***	0.02	0.07***	0.05**	0.03
(SE)	(0.03)	(0.03)	(0.04)	(0.02)	(0.02)	(0.02)
Controls & FE	Preferred controls; FE as in Table 2 - Col (4)					
Observations	189,859					
Clusters	109,112					

Notes: Each column reports the coefficient on the corresponding readability index from a separate regression of 28-day mortality on the index, ln(length), and the preferred control set with the full fixed-effects stack. Entries are coefficients multiplied by 100 (percentage points); robust standard errors clustered at the patient level in parentheses.
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table B9: Alternative Readability Indices: 28-Day Readmission

	SMOG	Masked	Dale–Chall	Flesch–Kincaid	Gunning Fog	Coleman–Liau
Coefficient	-0.03	-0.06	0.14	-0.02	0.00	0.02
(SE)	(0.08)	(0.08)	(0.11)	(0.06)	(0.05)	(0.06)
Controls & FE	Preferred controls; FE as in Table 2 - Col (4)					
Observations	189,859					
Clusters	109,112					

Notes: Each column reports the coefficient on the indicated readability index from a separate regression of 28-day readmission on the index, ln(length), and the preferred control set with the full fixed-effects stack. Entries are coefficients multiplied by 100 (percentage points); robust standard errors clustered at the patient level in parentheses.

Table B10: Alternative Readability Indices: 28-Day ED Visit

	SMOG	Masked	Dale–Chall	Flesch–Kincaid	Gunning Fog	Coleman–Liau
Coefficient	-0.02	-0.04	0.05	-0.03	0.01	0.01
(SE)	(0.08)	(0.08)	(0.11)	(0.06)	(0.05)	(0.06)
Controls & FE	Preferred controls; FE as in Table 2 - Col (4)					
Observations	189,859					
Clusters	109,112					

Notes: Each column reports the coefficient on the indicated readability index from a separate regression of 28-day ED revisit on the index, ln(length), and the preferred control set with the full fixed-effects stack. Entries are coefficients multiplied by 100 (percentage points); robust standard errors clustered at the patient level in parentheses.

B.5 Heterogeneity Analyses using LLM-Extracted Variables

Figure B1: SMOG (lexical difficulty) and CICS (procedural complexity) highlight distinct informational burdens.

Red-flag symptoms

Low CICS (tier 1).
"Call your doctor for shortness of breath or other concerning symptoms."
Features: Single trigger; no thresholds; no exceptions.
SMOG decile (within task): 2

High CICS (tier 4–5).
"Return for chest pain; worsening cough/wheeze; vomiting you can't keep fluids down; signs of dehydration (dry mouth, fast heartbeat, dizziness on standing); black stools or blood in vomit; fever >101.5°F; any new concerning symptom. Avoid driving on narcotics; call for incision redness/swelling/drainage; staples out at follow-up; remove steri-strips after X days."
Features: Long multi-trigger list, numeric thresholds, exceptions.
SMOG decile (within task): 9

Medication changes

Low CICS (tier 1–2).
"Resume home meds; avoid aspirin/ibuprofen for one week; call if fever >101.5°F or severe pain."
Features: Single action + one prohibition; no sequencing or titration.
SMOG decile (within task): 3

High CICS (tier 4–5).
"Increase aspirin to 325 mg; start clopidogrel (3 months); stop omeprazole (interaction); switch to ranitidine; hold lorazepam; change metoprolol formulation; recheck INR in 3 days; continue enoxaparin until INR ≥ 2."
Features: Multiple starts/stops, interactions, monitoring thresholds.
SMOG decile (within task): 7

Table B11: Moderation of SMOG by High Procedural Complexity (CICS $\geq$ 4)

Task (subsample)	Interaction: SMOG $\times$ High CICS	Std. Err.	p-value
Weight (N = 18,661)	0.268	0.173	0.120
Low-salt (N = 2,383)	-0.418	0.966	0.666
Stop/hold med (N = 34,891)	-0.042	0.165	0.800
Red-flag (N = 82,515)	-0.058	0.074	0.439

Notes: Each row reports the coefficient on SMOG interacted with an indicator for CICS $\geq$ 4 in the corresponding task-present subsample. All models mirror the preferred specification and cluster at the patient level. Coefficients and standard errors multiplied by 100 to express percentage-point changes per one-grade increase in SMOG.

Table B12: Red-Flag Position and the SMOG–Mortality Association

Coefficient (red-flag present)	Estimate[a]	Std. Err.	p-value
SMOG main effect	0.097	0.058	0.092
Position: middle (vs. bottom)	0.466	0.911	0.609
Position: top (vs. bottom)	0.232	0.999	0.816
SMOG×middle	-0.045	0.079	0.565
SMOG×top	-0.018	0.090	0.843
Obs. (N)		82,515	

Notes: Estimates from a single model with SMOG interacted with indicators for the position of the first red-flag block (top/middle/bottom), plus the preferred controls and fixed effects; clustered at the patient level. Coefficients and standard errors multiplied by 100 (percentage points per one-grade increase in SMOG).

C Example of a Full Discharge Summary

Appendix Figure C1: Stylized Example of a Full Discharge Summary (Part 1/2)

Admission Date: XXXX-XX-XX **Discharge Date:** XXXX-XX-XX
Service: MEDICINE (CARDIOLOGY)

Allergies: No Known Allergies

Chief Complaint: Chest pain

Major Surgical or Invasive Procedure: Left heart catheterization with percutaneous coronary intervention (PCI).

History of Present Illness: Mr. ____ is a 68-year-old male with a history of hypertension, hyperlipidemia, and a 40-pack-year history of smoking who presented to an outside hospital with 3 hours of substernal, non-radiating chest pressure. An EKG at the outside hospital was concerning for an NSTEMI, with an initial troponin of 0.8 ng/mL. He was started on a heparin drip and transferred to BIDMC for urgent cardiac catheterization.

Brief Hospital Course: The patient was admitted to the Cardiology service for management of NSTEMI. He was taken for cardiac catheterization on the day of admission. The procedure revealed an 80% stenosis of the mid-left anterior descending (LAD) artery, which was successfully treated with the placement of a drug-eluting stent. His post-procedural course was uncomplicated. His chest pain resolved completely. He was started on dual anti-platelet therapy (Aspirin and Clopidogrel) and a high-intensity statin. An echocardiogram showed a normal ejection fraction of 60% with no regional wall motion abnormalities. He was monitored on telemetry overnight without any arrhythmias. He remained hemodynamically stable and was deemed ready for discharge on hospital day 2.

Discharge Disposition: Home

Discharge Diagnosis:

- Non-ST Elevation Myocardial Infarction (NSTEMI)
- Coronary Artery Disease s/p PCI with DES to LAD
- Hypertension
- Hyperlipidemia
- Tobacco Use Disorder

Appendix Figure C1: Stylized Example of a Full Discharge Summary (Part 2/2)

Discharge Medications:

1. Aspirin 81 mg tablet, take one tablet by mouth daily.
2. Clopidogrel (Plavix) 75 mg tablet, take one tablet by mouth daily.
3. Atorvastatin (Lipitor) 80 mg tablet, take one tablet by mouth every evening.
4. Lisinopril 10 mg tablet, take one tablet by mouth daily.
5. Metoprolol Succinate XL 25 mg tablet, take one tablet by mouth daily.

Discharge Instructions:
Dear Mr. ____,
You were admitted to the hospital because you had a heart attack. A procedure was done where we looked at the arteries of your heart and placed a small metal scaffold called a stent to open up a blockage.
It is extremely important that you take your new heart medications every day to prevent another heart attack or a clot from forming in your new stent. You must take both Aspirin 81mg and Clopidogrel 75mg every single day. Do not stop taking these medications for any reason without first talking to your heart doctor (cardiologist). Stopping these medications can be life-threatening.
You should avoid any heavy lifting (more than 10 pounds) for one week. You can shower but should not soak the area where the catheter was placed in your wrist or groin. Watch for any bleeding, swelling, or redness at this site.
Please call our office to make a follow-up appointment with the cardiology clinic in 2-4 weeks. You should also see your primary care doctor within 7 days of discharge. Please call 911 if you experience any new or returning chest pain.

Followup Instructions: ____