BioMedpha Volume: 2 Issue:1 6/2025/pp.149-195

A journal of Life, Health, Biomedical, Pharmaceutical and Chemical Sciences

Health Data Literacy as a Public Health Intervention: Improving the Public's Ability to Understand and Use Health Information

[1] Munachiso Angela Muoneke

[2] Chinedu Collins Okafor

[1] University of the Potomac Washington DC, USA.

Email address: mamuoneke@gmail.com

Orcid ID: https://orcid.org/0009-0007-4298-879X

[2] Email address: collins.okafor18@gmail.com

University of North Carolina, Chapel Hill Country: United States

Orcid ID: https://orcid.org/0009-0000-1960-0389

Abstract

Today, people receive health information mainly in numerical form: risk percentages, laboratory results, screening probabilities and online dashboards. The ability to understand and use these numbers, referred to here as health data literacy, is therefore a public health issue. Population evidence on who has trouble with health statistics, and on how this difficulty relates to everyday information tasks, is still limited. We conducted a secondary cross-sectional analysis of the Health Information National Trends Survey 6 (HINTS 6, 2022; n = 6,252 U.S. adults). The primary outcome was self-reported difficulty understanding medical statistics (hard or very hard). Design-based analyses used the final person weight and 50 jackknife replicate weights. We fitted survey-weighted logistic and proportional-odds models, calculated adjusted predicted probabilities and examined dose–response relationships with five information outcomes. Among 6,011 adults with a valid answer, 26.8% (95% CI 24.7–28.8) reported difficulty. The adjusted odds of difficulty increased by 42% for each step down in education (adjusted odds ratio [aOR] 1.42, 95% CI 1.21–1.66). Not using the Internet (aOR 1.66, 1.20–2.29) and fair or poor health (aOR 2.82, 2.03–3.92) were also independent correlates. Among adults with recent health care visits, those whose clinicians did not always explain things clearly had higher odds of difficulty (aOR 2.72, 2.22–3.32). Difficulty was associated with low confidence completing

medical forms (aOR 2.34), low confidence finding online health resources (aOR 2.60), difficulty understanding cancer information (aOR 2.75) and difficulty understanding patient-portal information (aOR 5.30). Portal-information difficulty rose from 2.6% to 44.4% across the four response levels. About one in four U.S. adults report difficulty understanding medical statistics. This difficulty is socially patterned and strongly associated with practical barriers involving forms, online information, cancer information and patient portals. The results identify priority groups and delivery channels for health data literacy interventions. Because the design is observational, these interventions require controlled evaluation.

Keywords: health data literacy; health literacy; numeracy; risk communication; patient portals; digital health equity; public health intervention; HINTS

Introduction

Health information has become numerical. A person who logs in to a patient portal today may see a cholesterol value flagged "H", an HbA1c of 6.6% or a biopsy report, sometimes before a clinician has explained it. Since U.S. rules began requiring the immediate release of electronic results, a growing share of test results has been viewed by patients before clinicians (Steitz et al., 2021). A survey of more than 8,000 portal users found that 95.7% preferred immediate access, but 16.5% of those with abnormal results reported more worry after reading their results without an explanation (Steitz et al., 2023). The COVID-19 pandemic showed the same pattern at population level. Daily case counts, test-positivity rates and vaccine efficacy percentages became part of everyday life, and the visual format of a single official chart was enough to change people's perceptions of their own risk (Padilla et al., 2022).

Numbers can mislead as easily as they inform. A medicine that "halves the risk" of stroke sounds impressive, but if the absolute risk falls from 2 in 1,000 to 1 in 1,000, the practical benefit is small. Best-practice guidance therefore recommends absolute rather than relative risks, consistent denominators and fixed time frames, because inconsistent formats can bias perception and steer decisions (Bonner et al., 2021). Whether a person can see through such framing depends on their own skills and also on how clearly organisations present the information.

Healthy People 2030 recognises this shared responsibility. It defines personal health literacy as the ability to find, understand and use information for health decisions, and organisational health

literacy as the degree to which organisations enable people to do so (Brach & Harris, 2021; Santana et al., 2021). Health literacy is now understood as a social determinant of health that follows a social gradient and contributes to health disparities (Nutbeam & Lloyd, 2021; Schillinger, 2021). Low health literacy is associated with poorer health outcomes (Shahid et al., 2022). Richter and Krawczak (2024) argue that people need a specific form of health data literacy when they must evaluate information generated from health data. This paper focuses on its numerical core: the ability to understand and use health statistics.

The problem is large and may be growing. In the 2023 Survey of Adult Skills, 34% of U.S. adults performed at or below Level 1 in numeracy, compared with 29% in 2017 (National Center for Education Statistics, 2024; Organisation for Economic Co-operation and Development, 2024). In 17 European countries, between one in four and three in four adults had limited general health literacy, and 38% found digital health information tasks difficult (HLS19 Consortium of the WHO Action Network M-POHL, 2021). At the same time, health systems are moving information online, which makes digital inclusion a determinant of health in its own right (Sieck et al., 2021; van Kessel et al., 2022).

Much numeracy research is experimental or clinical. It shows how presentation formats change understanding, but it says less about where difficulty is concentrated in the population and which everyday information tasks it affects. A public health response needs two kinds of evidence: who should be reached, and through which channels. This study provides that evidence using a national probability sample.

Study Aim and Research Questions

The aim was to estimate the prevalence and social distribution of self-reported difficulty understanding medical statistics among U.S. adults, and to examine how this difficulty relates to clinician communication and to the use of health information. The study addresses four questions. RQ1: What proportion of adults report difficulty understanding medical statistics? RQ2: Which demographic, digital and health characteristics are independently associated with this difficulty? RQ3: Is this difficulty associated, in a graded manner, with confidence and comprehension in common health information tasks, including forms, online resources, cancer information and

patient portals? RQ4: What intervention priorities follow from these patterns? By linking self-reported difficulty understanding medical statistics to real-world information tasks in a national sample, the study supports targeted rather than generic health data literacy interventions.

Literature Review

From Health Literacy to Health Data Literacy

The concept of health literacy has shifted from an individual skill to a shared attribute of people, organisations and systems. Nutbeam and Lloyd (2021) describe it as a modifiable social determinant that can be strengthened through education and better communication. The Healthy People 2030 definitions clearly state that organisations are responsible for making information and services easy to find, understand and use (Brach & Harris, 2021; Santana et al., 2021). Sørensen et al. (2021) extend this idea to whole systems and describe how policies, workforce and services can be organised to be health literate. During the pandemic, health literacy was described as a 'social vaccine' that helps people follow official guidance and resist the infodemic (Okan et al., 2023). The World Health Organization (2022) also presents the development of health literacy as a strategy for preventing and controlling noncommunicable diseases.

Health data literacy is a narrower concept that focuses on information expressed as data. Richter and Krawczak (2024) define it as the capacity to find, understand and evaluate information about data-intensive health research and to act on it in line with one's values. In daily life, the most common form of health data is numerical: probabilities, rates, test values and charts. The ability to understand such numbers is therefore the foundation of health data literacy, and it is the dimension examined in this study.

Numeracy and the Communication of Health Statistics

Numeracy has subjective and objective dimensions. The Centers for Disease Control and Prevention (CDC, 2025) note that people may lack numerical skills, lack confidence in those skills, or both. Both dimensions matter. In two population surveys, Benda et al. (2021) found that lower scores on a three-item subjective numeracy scale and lower scores on an objective numeracy assessment were each associated with poorer self-rated health, with similar effect sizes. That study

examined a multi-item scale and does not establish the validity of single-item measures. Experimental research shows that presentation strongly influences understanding. Bonner et al. (2021) recommend presenting absolute numbers over a set time period with a clear denominator, using consistent formats across options and interpreting numbers for readers with different numeracy levels. Freeman et al. (2021) tested COVID-19 risk displays with 5,520 adults and recommended percentages supported by a contrasting frequency format and visual scales. Communicating uncertainty and the quality of evidence did not undermine trust in two randomised trials (Schneider et al., 2021), and balanced messages were judged at least as trustworthy as persuasive ones (Kerr et al., 2022). Padilla et al. (2022) found that cumulative and incident-scale charts of the same mortality data produced different risk perceptions. In a systematic review, Nan et al. (2022) identified literacy, numeracy, analytical thinking and trust in science as factors that protect against health misinformation.

Digital Health Information, Patient Portals and Unequal Access

Online sources are now a routine starting point for health information (Wang et al., 2021). Digital health literacy is linked to better self-management and engagement, and education and social support are effective ways to strengthen it (Arias López et al., 2023). However, it is unequally distributed. In a recent meta-analysis, Zhong et al. (2025) reported that education, income and Internet use were each positively associated with eHealth literacy. Digital tools have not yet reduced health inequities and may widen them for older, poorer and less educated groups (Lyles et al., 2023; Yao et al., 2022). Older adults face specific barriers to digital engagement (Kebede et al., 2022). Patients with limited English proficiency are less likely to use video telehealth (Rodriguez et al., 2021). Patient portals concentrate these challenges. Using HINTS 6, Maini et al. (2025) found that 61.3% of adults had accessed a portal in the past year and that offers and encouragement from providers were strong predictors of use. Richwine et al. (2023) showed that provider encouragement was associated with smaller disparities in portal access. Among hospital inpatients, lower eHealth literacy was associated with lower awareness and use of patient portals (Deshpande et al., 2023). Portals have widened access to data, but they do not by themselves ensure that people can interpret it.

Interventions to Improve Understanding and Use of Health Information

Evidence supports interventions at both the individual and the organisational level. The teach-back method, in which patients explain information back in their own words, is a core tool in the Health Literacy Universal Precautions Toolkit (Agency for Healthcare Research and Quality [AHRQ], 2024a, 2024b). In a systematic review, five of six teach-back studies reported significant reductions in hospital readmission (Mashhadi et al., 2021). Digital health literacy training for older adults improved eHealth literacy scores (standardised mean difference 1.15; Dong et al., 2023), although the included studies were heterogeneous. Clinician communication is also a modifiable resource: in HINTS data, patient-centred communication was consistently associated with greater trust in clinicians as an information source (Asan et al., 2021). Organisational approaches simplify materials, forms and processes so that fewer demands are placed on individual skills (CDC, 2024; Sørensen et al., 2021).

Conceptual Framework and Research Gap

Figure 1 summarises the framework that guided this study. Structural and social conditions shape the capacity to understand health statistics. In this study, that capacity is represented by a single measured indicator: self-reported difficulty understanding medical statistics. This capacity, in turn, affects capability in concrete information tasks. Interventions can act on the capacity itself, through education and support, or on the information environment, through clearer communication and simpler systems. Three gaps remain. First, few studies estimate the national prevalence of self-reported difficulty understanding medical statistics using design-based methods. Second, the link between this difficulty and specific everyday tasks, such as portals and forms, has rarely been tested with graded exposure measures. Third, the role of clinician communication as a modifiable correlate has received little attention in numeracy research. The present study addresses all three.

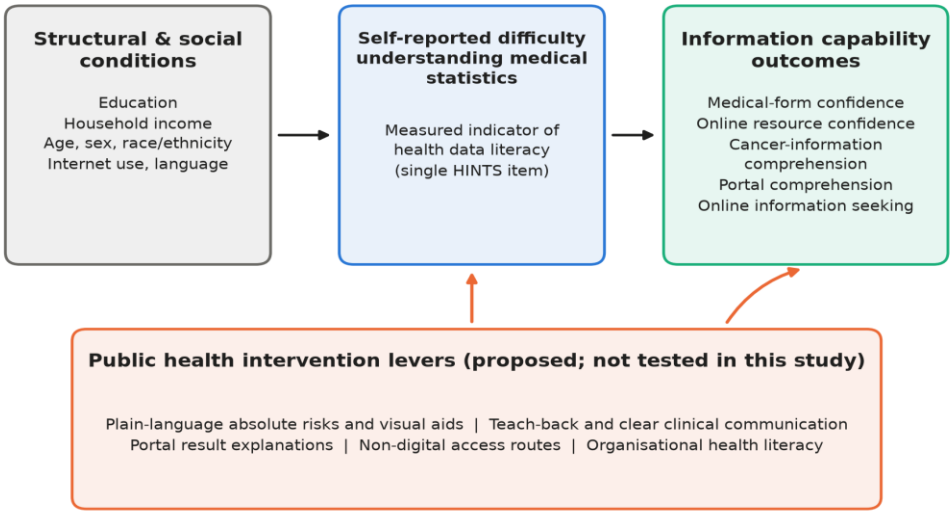


Figure 1. Conceptual framework. The central box shows the measured indicator, self-reported difficulty understanding medical statistics; arrows show hypothesised pathways, which this cross-sectional study tests only as associations.

Methodology

Research Design and Data Source

We used a quantitative, cross-sectional design based on secondary analysis of the Health Information National Trends Survey 6 (HINTS 6), conducted by the U.S. National Cancer Institute (NCI). HINTS 6 sampled civilian, non-institutionalised adults aged 18 years or older through address-based sampling. Questionnaires were completed on paper or online between 7 March and 8 November 2022. The public-use file contains 6,252 respondents, and the household response rate was 28.1% (NCI, 2023). We used the release updated in August 2025 (NCI, 2025).

This design was chosen for four reasons. First, HINTS 6 is a national probability sample with replicate weights, so it supports population-level estimates of prevalence and social distribution, which is what RQ1 and RQ2 require. Second, it is unusual in containing, in one instrument, a numeracy item alongside measures of forms, online information, cancer information, portal comprehension and clinician communication, which RQ3 requires. Third, it is publicly available and fully documented, so every estimate can be reproduced. Fourth, a cross-sectional design is

appropriate for describing burden and identifying targets, although it cannot establish causal order (Savitz & Wellenius, 2023). Other designs were considered and rejected. A new primary survey would have been costly and could not match the national coverage. The Survey of Adult Skills measures objective numeracy but contains no health information tasks. Qualitative methods would explain experiences but could not estimate prevalence. An intervention trial is the right next step, but it requires the targeting evidence that this study provides.

Measures

Table 1 lists all study variables. The primary outcome, self-reported difficulty understanding medical statistics, came from the question "How easy or hard is it for you to understand medical statistics?" Answers of hard or very hard were coded as difficulty. This single item records self-reported difficulty understanding medical statistics; it is not a validated numeracy scale or an objective test of numerical skill. It was used because it is the only HINTS 6 item that asks directly about understanding health statistics and because it is available alongside the information outcomes of interest. The full four-level response was retained for ordinal and dose–response analyses. Five information outcomes were defined using the official HINTS routing rules, so each applies only to respondents eligible for that question. Clinician communication was measured among respondents with at least one health care visit in the past year, using two items: how often clinicians explained things in a way the respondent could understand, and how often they helped the respondent deal with feelings of uncertainty about their health. Negative HINTS codes, which indicate missing, invalid or inapplicable answers, were treated as missing and never as a negative response.

Table 1. *Study variables, coding and eligible populations*

Construct	HINTS 6 item	Coding used in this study	Eligible population
Primary outcome			
Self-reported difficulty understanding medical statistics	UndMedicalStats	Hard/very hard = 1; easy/very easy = 0. Four levels kept for ordinal and gradient models	All respondents

Information capability outcomes			
Low confidence completing medical forms	ConfidentMedForms	A little/not at all = 1	All respondents
Low confidence finding helpful online health resources	ConfidentInternetHealth	A little/not at all confident = 1	All respondents
Cancer information hard to understand	CancerTooHardUnderstand	Strongly/somewhat agree = 1	Cancer-information seekers
Sought health information online (past 12 months)	Electronic2_HealthInfo	Yes = 1	Internet users
Portal information hard to understand	UnderstandOnlineMedRec2	Somewhat/very difficult = 1	Portal users (past 12 months)
Clinical communication			
Clinicians explained things clearly	ExplainedClearly	Not always (usually/sometimes/never) = 1	≥1 visit in past 12 months
Clinicians helped with uncertainty	HelpUncertainty	Not always = 1	≥1 visit in past 12 months
Covariates			
Demographic	Age, EducA, HHInc, BirthGender, RaceEthn5	Age 4 groups; education 4; household income 5; sex at birth 2; race/ethnicity 5	All respondents
Digital access	UseInternet	Does not use the Internet = 1	All respondents
Health and care context	GeneralHealth, HealthInsurance2,	Self-rated health 3 groups; uninsured; non-	All respondents

	RUC2013, LANGUAGE_FLAG, FreqGoProvider	metropolitan; Spanish questionnaire; visits 0/1–2/3–4/5+	

Note. Design variables: PERSON_FINWT0 (full-sample weight) and PERSON_FINWT1–50 (jackknife replicate weights). Exact wording and response labels follow the HINTS 6 codebook (NCI, 2025).

Statistical Analysis

All estimates were design-based. Point estimates used the full-sample weight. Variances used the 50 jackknife replicate weights with a multiplier of 0.98, as recommended by NCI (2023), with 49 degrees of freedom for t-based confidence intervals. We used design-based methods rather than unweighted models because HINTS oversamples some groups and adjusts for nonresponse; unweighted estimates would not represent the U.S. population, and naive standard errors would be too small.

Analysis proceeded in five steps. (1) We estimated weighted prevalence overall and by subgroup. (2) A survey-weighted logistic regression (the main model) included age, education, income, sex at birth, race/ethnicity and Internet use. Logistic regression was selected because the policy-relevant outcome is a binary state (reporting difficulty or not), and because it allows adjustment for several correlated social factors at once. Design-based Wald F tests assessed each factor as a whole. A linear education score tested for a gradient. Because odds ratios can overstate differences when an outcome is common, we also present adjusted predicted probabilities obtained by marginal standardisation. An extended model added self-rated health, insurance, rurality and questionnaire language. An education × Internet interaction tested whether the education association differed by Internet use. (3) A survey-weighted proportional-odds model used all four response levels, to check that results did not depend on how the outcome was dichotomised. The proportional-odds (parallel-lines) assumption was assessed with a design-based Brant-type test: separate survey-weighted logistic models were fitted to the same 5,214 complete cases for each of the three cumulative splits of the four-level response (4,292, 1,331 and 145 events, respectively), and equality of the coefficients across splits was tested with jackknife Wald F tests, both globally and for each covariate. (4) Two models examined clinician communication among respondents

with recent visits, adjusting for demographics, Internet use and health status. (5) For each information outcome, Model A adjusted for demographics (and Internet use where applicable), and Model B added health status, insurance, rurality and visit frequency. A gradient model entered the four-level response for self-reported difficulty understanding medical statistics as a linear term, with Model A covariates, to test for a dose–response relationship. Sample sizes for Model A, Model B and the gradient models are reported in Table 5. Benjamini–Hochberg q-values controlled the false discovery rate across the five outcomes.

Missing and inapplicable answers were excluded from each analysis rather than imputed. Following current guidance on reporting missing data (Carpenter & Smuk, 2021; Lee et al., 2021), we compared difficulty among complete and excluded cases, refitted the main model without income (the covariate with the most missing values), used “very hard” alone as an alternative outcome, and calculated bounds for missing primary responses. The original descriptive and primary analyses were executed in Python 3.12.14, as recorded in the execution log supplied with the analysis package, and their coefficients were checked against an independent optimiser. The extended analyses shown in Appendix A were run in Python 3.13.16 and re-run in Python 3.12.13 (NumPy 2.5.4, pandas 3.0.6, SciPy 1.18.1); all estimates were identical across Python versions.

Ethical Considerations and Reproducibility

HINTS 6 public-use data are de-identified and freely available from NCI. This study used only these publicly available data; no identifiable private information was accessed and there was no interaction with participants. Under the U.S. Common Rule (45 CFR 46.102), analysis of such data does not meet the definition of human subjects research, so institutional review board approval was not required. The complete analysis code, which reproduces every number, table and figure in this article from the original data file, is provided as supplementary material and shown in Appendices A and B.

Results

Prevalence and Distribution of Difficulty

Of 6,252 respondents, 6,011 answered the medical-statistics question. Most found medical statistics easy (56.7%) or very easy (16.6%), but 24.2% found them hard and 2.6% very hard. The weighted prevalence of difficulty was therefore 26.8% (95% CI 24.7–28.8), equal to roughly one in four U.S. adults. Table 2 and Figure 2 show the unadjusted distribution. The steepest gradient was by education: prevalence fell from 41.4% among adults without a high school diploma to 18.0% among college graduates. Difficulty was also high among adults who did not use the Internet (42.3% vs 24.2%), those in fair or poor health (44.2% vs 18.2% for excellent or very good health), uninsured adults (36.6%), lower-income households (36.3% below $20,000) and respondents who completed the Spanish questionnaire (43.9%).

Table 2. *Sample characteristics and weighted prevalence of self-reported difficulty understanding medical statistics (N = 6,011)*

Characteristic	Unweighted n	Weighted % of sample	Difficulty, weighted % (95% CI)
Age			
18–34 years	879	25.4	27.1 (23.0–31.3)
35–49 years	1,175	25.0	23.8 (19.7–27.8)
50–64 years	1,727	27.8	25.6 (22.5–28.8)
65 years or older	2,146	21.8	30.9 (27.9–34.0)
Education			
Less than high school	379	6.8	41.4 (30.5–52.2)
High school graduate	1,051	21.6	36.5 (31.0–42.0)
Some college	1,656	38.9	24.5 (21.0–27.9)
College graduate or more	2,709	32.8	18.0 (15.9–20.2)

Annual household income			
Less than $20,000	941	14.3	36.3 (29.9–42.8)
$20,000–$34,999	722	11.0	29.8 (22.8–36.7)
$35,000–$49,999	725	11.8	29.7 (23.8–35.6)
$50,000–$74,999	928	18.2	23.0 (19.1–27.0)
$75,000 or more	2,157	44.6	21.7 (18.2–25.3)
Sex at birth			
Male	2,287	49.3	27.6 (24.4–30.7)
Female	3,502	50.7	25.7 (23.2–28.3)
Race and ethnicity			
Non-Hispanic White	3,184	61.3	23.1 (20.6–25.5)
Non-Hispanic Black	877	10.9	27.7 (23.0–32.3)
Hispanic	993	17.0	32.6 (26.1–39.0)
Non-Hispanic Asian	288	5.7	31.7 (20.1–43.3)
Non-Hispanic Other	184	5.1	34.0 (17.0–51.0)
Internet use			
Uses the Internet	4,980	85.8	24.2 (22.0–26.4)
Does not use the Internet	1,029	14.2	42.3 (37.0–47.5)
Self-rated health			
Excellent or very good	2,664	46.2	18.2 (15.3–21.1)
Good	2,234	36.7	29.4 (25.8–33.0)
Fair or poor	1,079	17.1	44.2 (38.4–50.1)
Health insurance			

Insured	5,478	89.4	25.4 (23.2–27.7)
Uninsured	496	10.6	36.6 (27.8–45.3)
Questionnaire language			
English	5,875	97.9	26.4 (24.3–28.5)
Spanish	136	2.1	43.9 (31.1–56.7)

Note. Overall prevalence 26.8% (95% CI 24.7–28.8). Each characteristic uses its own valid denominator; missing values: age 84, education 216, income 538, sex at birth 222, race/ethnicity 485, Internet use 2, health 34, insurance 37. Percentages are unadjusted.

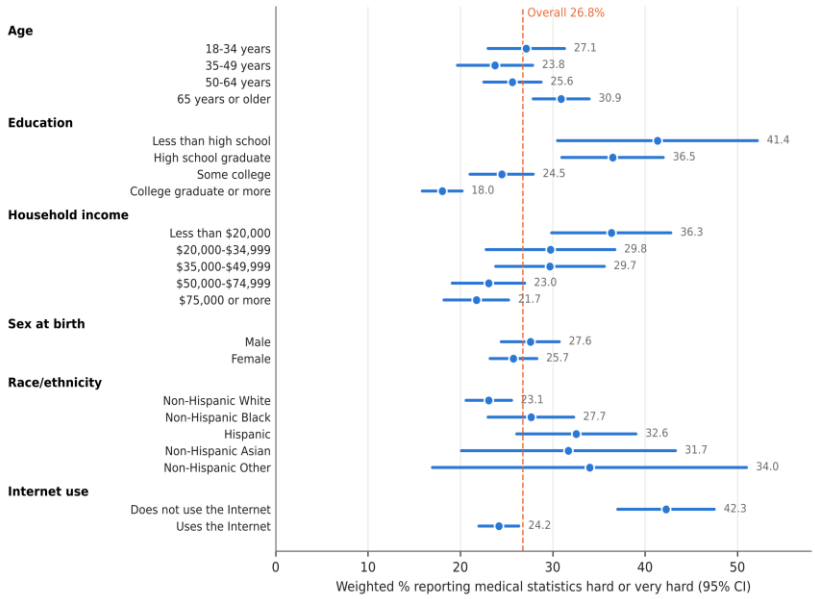


Figure 2. Weighted prevalence of self-reported difficulty understanding medical statistics by population group, with 95% confidence intervals. The dashed line marks the national estimate (26.8%).

Independent Correlates of Difficulty

Table 3 and Figure 3 present the adjusted models. In the main model (n = 5,214; 1,331 with difficulty), education was the strongest factor overall ($F(3, 47) = 8.66$, $p < .001$). Compared with college graduates, the odds of difficulty were higher among high school graduates (aOR 2.43, 95%

CI 1.72–3.44), adults without a high school diploma (aOR 2.11, 1.14–3.89) and those with some college (aOR 1.35, 1.03–1.77). Each step down the education ladder increased the odds by 42% (aOR 1.42, 1.21–1.66). Not using the Internet was independently associated with difficulty (aOR 1.66, 1.20–2.29). In the adjusted model, age, income and sex at birth were not significantly associated with difficulty. Income contrasts were not statistically significant after adjustment; this analysis did not establish mediation. The non-Hispanic Asian contrast exceeded 1 (aOR 1.78, 1.03–3.08), but race and ethnicity as a whole were not significant ($F(4, 46) = 1.65$, $p = .179$).

The extended model showed that self-rated health was a strong additional correlate ($F(2, 48) = 19.72$, $p < .001$). Adults in fair or poor health had nearly three times the adjusted odds of difficulty (aOR 2.82, 2.03–3.92), and those in good health had higher odds (aOR 1.61, 1.19–2.19) than those in excellent or very good health. The high school education (aOR 2.22) and Internet-use (aOR 1.58) associations remained. Insurance, rurality and questionnaire language were not independently associated once other factors were included. The proportional-odds model, which used all four response levels, showed the same pattern for education and Internet use (Table 3). The global test of the proportional-odds assumption did not show a statistically significant violation ($F(32, 18) = 1.90$, $p = .076$), and the tests for education ($F(6, 44) = 0.86$, $p = .532$) and Internet use ($F(2, 48) = 0.98$, $p = .382$) did not indicate departures. However, the assumption was not met for household income ($F(8, 42) = 2.58$, $p = .022$), so income estimates from the ordinal model should be interpreted with caution.

Table 3. *Adjusted odds ratios (95% CI) for self-reported difficulty understanding medical statistics*

Characteristic (reference)	Main model	Extended model	Ordinal model
Age (18–34 years)			
35–49 years	0.88 (0.64–1.22)	0.85 (0.61–1.18)	0.89 (0.67–1.17)
50–64 years	0.93 (0.67–1.30)	0.88 (0.62–1.23)	0.96 (0.73–1.26)
65 years or older	1.15 (0.80–1.65)	1.13 (0.80–1.59)	1.11 (0.82–1.50)
Education (college graduate or more)			
Less than high school	2.11 (1.14–3.89)	1.73 (0.90–3.32)	2.07 (1.08–3.97)

High school graduate	2.43 (1.72–3.44)	2.22 (1.57–3.15)	2.07 (1.46–2.95)
Some college	1.35 (1.03–1.77)	1.23 (0.93–1.63)	1.29 (1.03–1.61)
Household income ($75,000 or more)			
Less than $20,000	1.20 (0.79–1.82)	0.98 (0.63–1.52)	1.25 (0.87–1.79)
$20,000–$34,999	0.95 (0.61–1.47)	0.84 (0.55–1.28)	0.91 (0.60–1.40)
$35,000–$49,999	1.12 (0.73–1.72)	0.98 (0.65–1.47)	1.33 (0.93–1.90)
$50,000–$74,999	0.83 (0.61–1.14)	0.76 (0.55–1.03)	0.91 (0.70–1.17)
Sex at birth (male)			
Female	0.95 (0.76–1.19)	0.96 (0.77–1.21)	1.01 (0.84–1.21)
Race and ethnicity (non-Hispanic White)			
Non-Hispanic Black	1.20 (0.88–1.64)	1.11 (0.81–1.52)	1.21 (0.95–1.56)
Hispanic	1.30 (0.89–1.88)	1.14 (0.80–1.63)	1.29 (0.89–1.87)
Non-Hispanic Asian	1.78 (1.03–3.08)	1.57 (0.90–2.74)	2.02 (1.33–3.09)
Non-Hispanic Other	1.21 (0.64–2.29)	1.06 (0.55–2.05)	1.38 (0.90–2.12)
Internet use (uses the Internet)			
Does not use the Internet	1.66 (1.20–2.29)	1.58 (1.13–2.21)	1.61 (1.15–2.26)
Health context (extended model only)			
Good health (vs excellent/very good)	–	1.61 (1.19–2.19)	–
Fair or poor health	–	2.82 (2.03–3.92)	–
Uninsured	–	1.49 (0.96–2.33)	–
Non-metropolitan residence	–	1.03 (0.80–1.32)	–
Spanish-language questionnaire	–	0.99 (0.46–2.14)	–

Note. Main and ordinal models n = 5,214; extended model n = 5,163. Bold indicates 95% CI excluding 1. Ordinal model: survey-weighted proportional-odds model of the four response levels (higher = harder). Joint Wald tests (main model): education F(3, 47) = 8.66, p < .001; Internet use F(1, 49) = 9.92, p = .003; race/ethnicity F(4, 46) = 1.65, p = .179; income F(4, 46) = 1.45, p = .234; age F(3, 47) = 1.47, p = .234. Linear education trend: aOR 1.42 (1.21–1.66) per step lower. Proportional-odds assumption (Brant-type jackknife Wald tests of equal coefficients across the three cumulative splits; n = 5,214): global F(32, 18) = 1.90, p = .076; education F(6, 44) = 0.86, p = .532; Internet use F(2, 48) = 0.98, p = .382; age F(6, 44) = 0.73, p = .627; sex at birth F(2, 48) = 1.28, p = .287; race/ethnicity F(8, 42) = 1.21, p = .315; household income F(8, 42) = 2.58, p = .022 (assumption not met; interpret ordinal income estimates with caution).

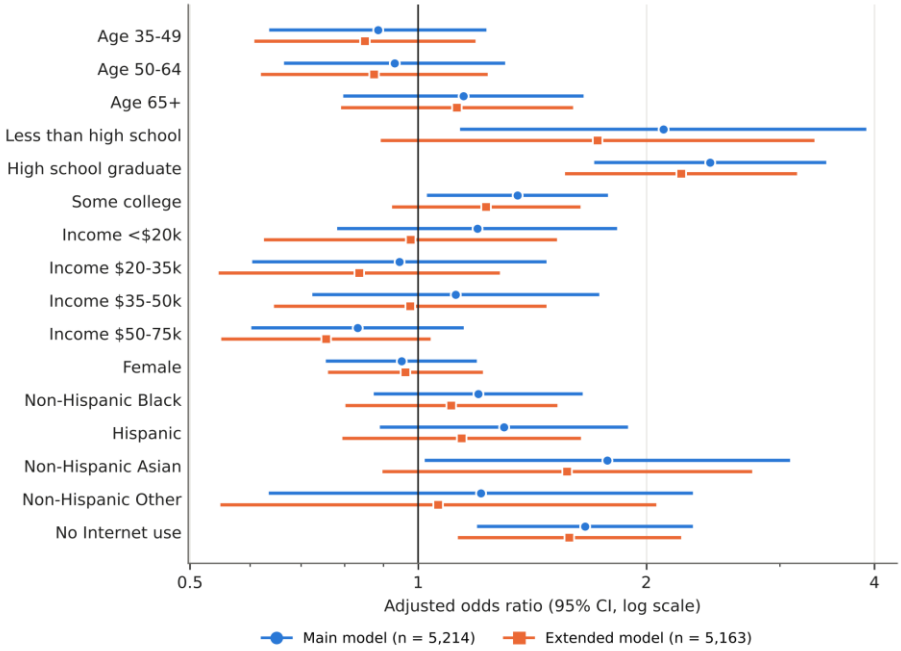


Figure 3. Adjusted odds ratios for self-reported difficulty understanding medical statistics in the main and extended models. Horizontal lines are 95% confidence intervals; the vertical line marks no association.

Adjusted predicted probabilities translate these odds ratios into absolute terms (Figure 4). After marginal standardisation over respondents' observed covariates, the predicted probability of difficulty was 35.8% for high school graduates and 18.9% for college graduates, and 34.0% for

non-users of the Internet compared with 24.0% for users. Point estimates were higher for non-users of the Internet at every education level, although several confidence intervals overlapped. The education × Internet interaction was not significant (F(3, 47) = 0.75, p = .529). There was no statistically significant evidence that the education association differed by Internet use.

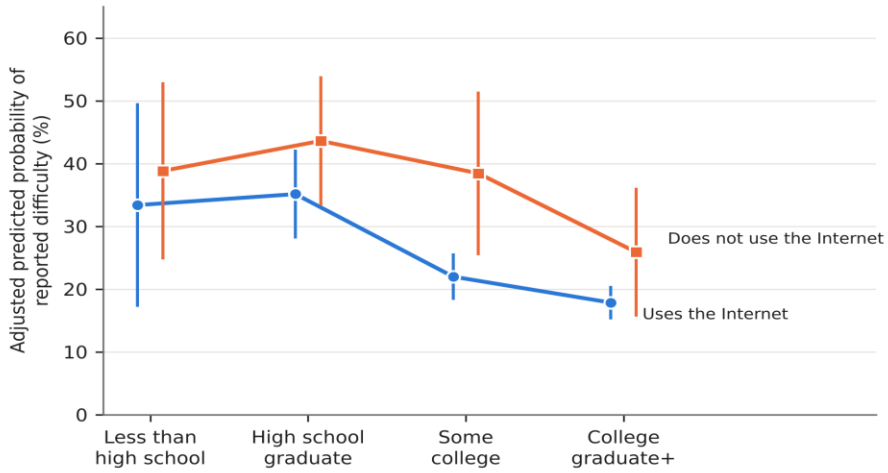


Figure 4. Adjusted predicted probability of self-reported difficulty understanding medical statistics by education and Internet use (marginal standardisation from the model including the education × Internet interaction; 95% CI). Interaction test: F(3, 47) = 0.75, p = .529.

Clinical Communication and Difficulty

Among adults with at least one health care visit in the past year, 41.1% reported that clinicians did not always explain things in a way they could understand. In this group, 36.9% reported difficulty with medical statistics, compared with 17.4% of those whose clinicians always explained clearly (Table 4). After adjustment for demographics, Internet use and health status, the odds of difficulty were 2.72 times higher (95% CI 2.22–3.32). A similar association was found for clinicians not always helping patients deal with uncertainty (aOR 2.34, 1.82–3.02).

Table 4. *Clinician communication and self-reported difficulty understanding medical statistics among adults with a health care visit in the past 12 months*

Clinician communication item	Always: difficulty % (95% CI)	Not always: difficulty % (95% CI)	Model n	aOR (95% CI)

Explained things in a way you could understand	17.4 (15.3–19.6)	36.9 (33.4–40.5)	4,571	2.72 (2.22–3.32)
Helped you deal with feelings of uncertainty	17.0 (14.5–19.5)	31.0 (28.0–34.0)	4,547	2.34 (1.82–3.02)

Note. Not always = usually, sometimes or never. Models adjust for age, education, income, sex at birth, race/ethnicity, Internet use and self-rated health. $p < .001$ for both associations. The direction of the association cannot be established from cross-sectional data.

Difficulty and Health Information Capability

Self-reported difficulty understanding medical statistics was strongly associated with four of the five information outcomes (Table 5; Figure 6). Compared with adults who found statistics easy, those who found them hard had higher adjusted odds of low confidence completing medical forms (aOR 2.34, 1.62–3.39), low confidence finding helpful online health resources (aOR 2.60, 1.87–3.62), finding cancer information hard to understand (aOR 2.75, 1.91–3.95) and finding portal information hard to understand (aOR 5.30, 3.61–7.78). All four associations had $q < .001$ and changed little after further adjustment for health status, insurance, rurality and visit frequency (Model B).

The associations followed a clear dose–response pattern (Figure 5). Across the four response levels, from very easy to very hard, portal difficulty rose from 2.6% to 44.4%, low form confidence from 7.3% to 43.0%, low online confidence from 8.4% to 55.3% and cancer-information difficulty from 28.7% to 80.0%. Each one-level increase in difficulty raised the odds of portal difficulty almost fourfold (aOR 3.80, 2.89–4.99). The pattern for online information seeking was different. In the binary comparison, the association was inconclusive (aOR 0.76, 0.54–1.08; $q = .126$). Across the four levels, however, seeking fell from 89.0% to 69.2%, and the gradient model showed a modest decline per level (aOR 0.77, 0.62–0.94; $p = .012$). Most Internet users who reported difficulty understanding medical statistics still looked for health information online, but they were less confident in finding helpful resources and, among portal users, more likely to find portal information hard to understand.

Table 5. *Associations between self-reported difficulty understanding medical statistics and health information outcomes*

Outcome (eligible group)	Easy %	Hard %	Model A aOR (95% CI)	Model B aOR (95% CI)	Gradient aOR per level (95% CI)
Low confidence with medical forms (all)	10.8	24.0	2.34 (1.62–3.39)	2.25 (1.55–3.27)	1.86 (1.38–2.52)
Low confidence finding online resources (all)	9.6	23.5	2.60 (1.87–3.62)	2.41 (1.79–3.23)	2.03 (1.63–2.52)
Cancer information hard to understand (seekers)	32.0	57.9	2.75 (1.91–3.95)	2.81 (1.91–4.15)	1.87 (1.48–2.35)
Sought health information online (Internet users)	85.8	80.4	0.76 (0.54–1.08)	0.71 (0.49–1.03)	0.77 (0.62–0.94)
Portal information hard to understand (portal users)	5.8	26.6	5.30 (3.61–7.78)	5.10 (3.41–7.63)	3.80 (2.89–4.99)

Note. Easy/Hard % = weighted prevalence of the outcome among adults finding medical statistics easy/very easy vs hard/very hard. Model A adjusts for age, education, income, sex at birth, race/ethnicity and (for forms, online confidence and cancer information) Internet use. Model B adds self-rated health, insurance, rurality and visit frequency. Model n, in row order: Model A 5,203; 5,138; 2,456; 4,420; 3,249. Model B 5,127; 5,059; 2,428; 4,366; 3,209. Gradient model 5,203; 5,138; 2,456; 4,420; 3,249 (same complete cases as Model A). Benjamini–Hochberg $q < .001$ for all Model A associations except online seeking ($q = .126$). The gradient model enters the four-level response (very easy = 1 to very hard = 4) as a linear term with Model A covariates.

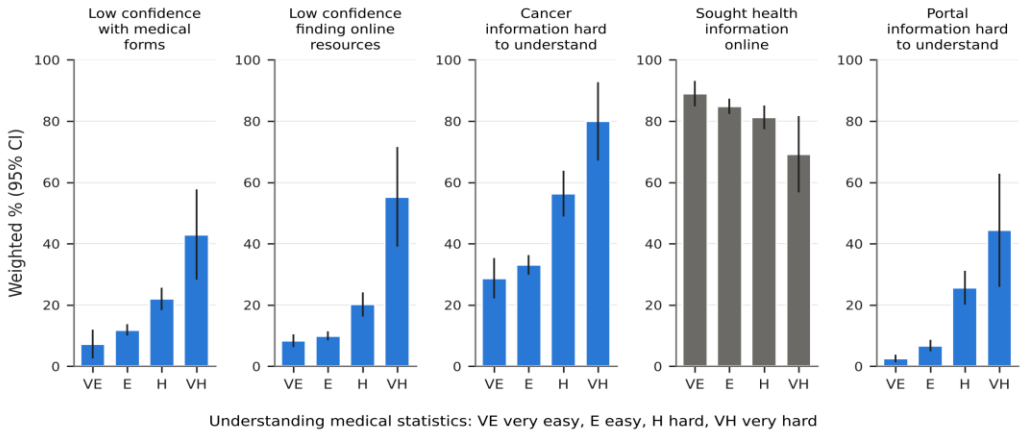


Figure 5. Weighted prevalence of each health information outcome across the four levels of self-reported difficulty understanding medical statistics (95% CI). VE = very easy, E = easy, H = hard, VH = very hard.

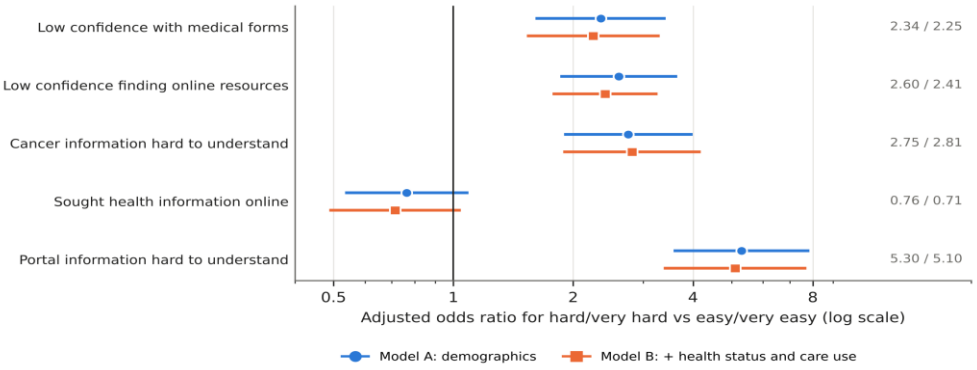


Figure 6. Adjusted odds ratios comparing adults who report finding medical statistics hard or very hard with those who find them easy, for five information outcomes. Model A adjusts for demographics; Model B also adjusts for health status and care use. Values on the right are Model A / Model B estimates.

Sensitivity and Missing-Data Analyses

Table 6 summarises the robustness checks. Removing income increased the analytic sample to 5,467, and the education and Internet-use associations were unchanged or slightly stronger. Respondents excluded from the main model because of missing covariates had higher prevalence of difficulty (37.4%) than those included (25.4%), so complete-case estimates may understate the

burden among the most disadvantaged. Assigning all 241 missing primary responses to either category gave bounds of 25.6% to 29.9%. When the outcome was restricted to very hard answers (145 events), income rather than education was the most prominent correlate and the estimates were imprecise. The most extreme difficulty may therefore follow a different, more material, social pattern than difficulty overall.

Table 6. *Sensitivity and missing-data analyses*

Analysis	Sample	Key result
Primary prevalence	n = 6,011	26.8% (24.7–28.8)
Very hard only	n = 6,011	2.6% (1.8–3.3)
Bounds for missing primary responses	241 missing	25.6% to 29.9% (assumption-based bounds, not CIs)
Complete cases vs excluded cases	5,214 vs 797	25.4% (23.2–27.6) vs 37.4% (30.1–44.7)
Main model omitting income	n = 5,467	High school vs college aOR 2.39 (1.71–3.34); no Internet aOR 1.73 (1.28–2.34)
Proportional-odds model (four levels)	n = 5,214	Same direction and significance for education and Internet use (Table 3); proportional-odds test: global p = .076, income p = .022
Education × Internet interaction	n = 5,214	F(3, 47) = 0.75, p = .529
Very-hard-only logistic model	n = 5,214; 145 events	Income <$75,000 contrasts aOR 3.43–6.49; education CIs include 1; sparse data

Note. All estimates are survey-weighted with jackknife variance estimation.

Discussion

This national study shows that about one in four U.S. adults report difficulty understanding medical statistics. This difficulty is not evenly distributed. It is more common among adults with

less education, those who do not use the Internet and those in poorer health. It is also strongly associated with the information tasks that modern health care depends on: completing forms, finding reliable online information, understanding cancer information and interpreting portal results. Three findings are particularly relevant to public health practice.

First, after adjustment, the social pattern was associated mainly with education and Internet use rather than with age or income. The strong education gradient is consistent with evidence that health literacy follows social gradients and acts as a social determinant of health (Nutbeam & Lloyd, 2021; Schillinger, 2021). It also matches European evidence that limited health literacy is concentrated in socially disadvantaged groups (HLS19 Consortium of the WHO Action Network M-POHL, 2021). The independent association with not using the Internet supports the view that digital inclusion is a determinant of health (Sieck et al., 2021; van Kessel et al., 2022). There was no statistically significant evidence that the education association differed by Internet use. Income contrasts were not statistically significant after adjustment; this analysis did not establish mediation. Education, income and Internet use are each associated with eHealth literacy in meta-analytic evidence (Zhong et al., 2025), and the pathways linking them to self-reported difficulty understanding medical statistics require longitudinal study. When the outcome was restricted to very hard responses, income contrasts were larger, and the proportional-odds test indicated that the income association differed across response levels. This raises the possibility that the most severe difficulty is more closely linked to material deprivation, but the small number of events means that this requires confirmation. The higher odds among non-Hispanic Asian adults were not supported by the joint test for race and ethnicity. This result may reflect language and acculturation barriers in a heterogeneous group (Rodriguez et al., 2021) rather than a stable ethnic difference, and it should not be over-interpreted.

Second, both poorer health and less clear clinical communication were significantly associated with difficulty. Adults reporting fair or poor health had nearly three times the adjusted odds of reporting difficulty. This is a concern because people in poorer health often need to manage the most numerical information, and low health literacy is associated with poorer health outcomes (Shahid et al., 2022). Adults whose clinicians did not always explain things clearly had 2.7 times the adjusted odds of difficulty. The direction of this association cannot be determined. Unclear

explanations may leave patients less confident with numbers, or patients who find numbers difficult may perceive explanations as less clear. In either case, the clinical encounter is where difficulty understanding medical statistics becomes visible and where it can be addressed. This supports universal-precautions approaches, which assume that any patient may have difficulty, and confirming understanding through teach-back (AHRQ, 2024a, 2024b; Mashhadi et al., 2021). It is also consistent with HINTS evidence that patient-centred communication builds trust in clinicians as a source of information (Asan et al., 2021).

Third, the strongest association was with patient portals. Among portal users, those who reported difficulty understanding medical statistics had five times the adjusted odds of finding portal information hard to understand, and portal difficulty was more than fifteen times higher at the hardest response level than at the easiest. This finding is especially important because results are increasingly released to patients before clinicians have reviewed them (Steitz et al., 2021). Most patients want this access, but a substantial minority with abnormal results report increased worry (Steitz et al., 2023). Portals deliver data, but without explanation they place the full burden of interpretation on the patient. Provider offers and encouragement are associated with higher portal use and smaller disparities in access (Maini et al., 2025; Richwine et al., 2023), but our findings suggest that access is not the same as understanding. Lower eHealth literacy is associated with lower portal engagement (Deshpande et al., 2023), and digital health tools have not yet reduced health inequities (Lyles et al., 2023; Yao et al., 2022). The findings on online information seeking support this conclusion. Adults who reported difficulty understanding medical statistics still searched for health information online at high rates, but they were less confident in finding helpful resources. This combination may increase vulnerability to misinformation, because literacy and numeracy are among the factors that protect against it (Nan et al., 2022).

Overall, the findings support treating health data literacy as a target for multilevel public health action rather than as a fixed individual capacity. Table 7 links each finding to an intervention priority and to the supporting evidence. Interventions on the information environment are likely to have the widest reach. Using absolute risks, consistent denominators and simple visual formats reduces the skill required of every reader (Bonner et al., 2021; Freeman et al., 2021), and such transparent communication does not reduce trust (Kerr et al., 2022; Schneider et al., 2021). Skill-

building interventions, such as digital health literacy training, can improve capability among older adults (Dong et al., 2023), and education and social support have been effective in strengthening digital health literacy (Arias López et al., 2023). Organisational and system-level approaches can make these measures routine (Brach & Harris, 2021; Sørensen et al., 2021; World Health Organization, 2022). As the pandemic showed, health literacy can work as a social vaccine when it is integrated into public health strategy (Okan et al., 2023).

Table 7. *Evidence-informed priorities for health data literacy interventions derived from the findings*

Finding from this study	Intervention priority	Supporting evidence	Suggested evaluation outcome
Difficulty in 26.8% of adults; steep education gradient	Universal plain-number communication: absolute risks, fixed denominators, icon arrays	Bonner et al. (2021); Freeman et al. (2021)	Objective comprehension of risk information
Portal difficulty aOR 5.30; strongest gradient	Plain-language explanations attached to released results; assisted portal navigation	Steitz et al. (2023); Richwine et al. (2023)	Correct interpretation of real test results; worry
Unclear clinician explanations aOR 2.72	Teach-back and universal precautions in routine care	AHRQ (2024a, 2024b); Mashhadi et al. (2021)	Teach-back accuracy; patient-reported clarity
No Internet use aOR 1.66; no significant interaction with education	Non-digital routes (print, telephone, community navigators) alongside digital training	Sieck et al. (2021); Dong et al. (2023)	Reach among offline adults; eHealth literacy
Fair/poor health aOR 2.82	Numeracy support built into chronic-disease and cancer care pathways	Shahid et al. (2022); WHO (2022)	Self-management and decision quality

Low online confidence despite high seeking	Signposting to trusted sources; misinformation resilience	Nan et al. (2022); Okan et al. (2023)	Source evaluation skills

Note. Interventions are proposed on the basis of observed associations and published evidence; none was implemented or evaluated in this study.

The study has several strengths. It uses a national probability sample with design-based variance estimation. It links self-reported difficulty understanding medical statistics to five concrete information tasks and to clinician communication. It tests graded as well as binary associations, and its sensitivity analyses show where conclusions are robust and where they are not. The full analysis can be reproduced from public data. Two implications follow for research and practice. First, the HINTS medical-statistics item is short enough to serve as a routine surveillance indicator of self-reported difficulty understanding medical statistics, provided that its validity against objective measures is established; it could then track trends and the effects of policy changes, such as the immediate release of results. Second, the priorities in Table 7 should be tested in controlled designs. For example, a trial could compare standard portal result displays with displays that add a plain-language explanation, an absolute-risk statement and a simple visual, measuring objective comprehension and worry before and after exposure. Adults with less education, those who are offline and those in poorer health carry the greatest burden and should be deliberately recruited into such trials.

Limitations

Several limitations should be considered. The design is cross-sectional, so the associations cannot establish temporal order or causation, and the proposed interventions were not tested. Difficulty was measured with a single self-reported item. This item records self-reported difficulty understanding medical statistics rather than objective numerical skill, has not been validated as a stand-alone numeracy measure, and does not assess graph reading, uncertainty or the evaluation of evidence (CDC, 2025). The primary measure and most outcomes were self-reported, so shared response tendencies may have inflated some associations. The response rate was 28.1%. Weighting reduces some sources of nonresponse bias but cannot eliminate it. Respondents excluded because of missing covariates reported greater difficulty than included respondents, so the burden may be

underestimated. The cancer, online-seeking and portal outcomes apply only to their eligible groups. Residual confounding is possible because cognitive function, English proficiency and prior health experience were not measured directly. The clinician communication items were available only for adults with recent visits, so those findings do not apply to people outside regular care. Several subgroup comparisons were exploratory, and some estimates, particularly for the very-hard-only outcome and small racial and ethnic groups, were imprecise. The proportional-odds assumption was not met for household income, so ordinal-model income estimates should be interpreted with caution. Finally, the data describe U.S. adults in 2022, and patterns may differ in other health systems and in later years.

Conclusion

Understanding health numbers is a widespread challenge rather than a marginal one. About one in four U.S. adults report difficulty understanding medical statistics, and this difficulty is most common among adults with less education, those who are offline and those in poorer health. It is strongly associated with unclear clinical communication and with barriers to completing forms, using online information, understanding cancer information and, above all, understanding patient-portal information. Health data literacy should therefore be treated as a public health priority. The response should address both sides of the problem: building people's skills and making numerical information easier to understand at its source. Clear numerical communication, teach-back in clinical care, explained portal results and non-digital access routes are the most promising starting points. Controlled trials that measure objective comprehension are now needed to determine which approaches improve understanding and health decisions. Regular monitoring in national surveys would show whether these gaps are narrowing. As more health information reaches people as numbers, the ability to understand those numbers will increasingly determine who benefits from it.

Declarations

Data availability: HINTS 6 public-use data are available from the National Cancer Institute (https://hints.cancer.gov). The analysis code is provided as supplementary material and is reproduced in Appendix A.

Ethics statement: This study was a secondary analysis of publicly available, de-identified HINTS 6 data. No identifiable private information was accessed and there was no interaction with participants; under the U.S. Common Rule (45 CFR 46.102), the analysis does not constitute human subjects research, and institutional review board approval was therefore not required.

Funding: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. **Conflicts of interest:** The author declares no conflicts of interest. **Author contributions:** The author was responsible for conceptualisation, methodology, formal analysis, data curation, visualisation, writing – original draft, and writing – review and editing.

References

Agency for Healthcare Research and Quality. (2024a). *Health literacy universal precautions toolkit* (3rd ed.). https://www.ahrq.gov/health-literacy/improve/precautions/toolkit.html

Agency for Healthcare Research and Quality. (2024b). *Use the teach-back method: Tool 5*. Health literacy universal precautions toolkit (3rd ed.). https://www.ahrq.gov/health-literacy/improve/precautions/tool5.html

Arias López, M. del P., Ong, B. A., Borrat Frigola, X., Fernández, A. L., Hicklent, R. S., Obeles, A. J. T., Rocimo, A. M., & Celi, L. A. (2023). Digital literacy as a new determinant of health: A scoping review. *PLOS Digital Health, 2*(10), e0000279. https://doi.org/10.1371/journal.pdig.0000279

Asan, O., Yu, Z., & Crotty, B. H. (2021). How clinician-patient communication affects trust in health information sources: Temporal trends from a national cross-sectional survey. *PLOS ONE, 16*(2), e0247583. https://doi.org/10.1371/journal.pone.0247583

Benda, N. C., Yang, Z., Li, H., Zhang, T., & Ancker, J. S. (2021). Lower objectively and subjectively assessed numeracy are both associated with poorer self-rated health. *BMC Research Notes, 14*, 321. https://doi.org/10.1186/s13104-021-05737-y

Bonner, C., Trevena, L. J., Gaissmaier, W., Han, P. K. J., Okan, Y., Ozanne, E., Peters, E., Timmermans, D., & Zikmund-Fisher, B. J. (2021). Current best practice for presenting

probabilities in patient decision aids: Fundamental principles. *Medical Decision Making, 41*(7), 821–833. https://doi.org/10.1177/0272989X21996328

Brach, C., & Harris, L. M. (2021). Healthy People 2030 health literacy definition tells organizations: Make information and services easy to find, understand, and use. *Journal of General Internal Medicine, 36*(4), 1084–1085. https://doi.org/10.1007/s11606-020-06384-y

Carpenter, J. R., & Smuk, M. (2021). Missing data: A statistical framework for practice. *Biometrical Journal, 63*(5), 915–947. https://doi.org/10.1002/bimj.202000196

Centers for Disease Control and Prevention. (2024, October 16). *Understanding health literacy*. https://www.cdc.gov/health-literacy/php/about/understanding.html

Centers for Disease Control and Prevention. (2025, January 31). *Numeracy*. https://www.cdc.gov/health-literacy/php/research-summaries/numeracy.html

Deshpande, N., Arora, V. M., Vollbrecht, H., Meltzer, D. O., & Press, V. (2023). eHealth literacy and patient portal use and attitudes: Cross-sectional observational study. *JMIR Human Factors, 10*, e40105. https://doi.org/10.2196/40105

Dong, Q., Liu, T., Liu, R., Yang, H., & Liu, C. (2023). Effectiveness of digital health literacy interventions in older adults: Single-arm meta-analysis. *Journal of Medical Internet Research, 25*, e48166. https://doi.org/10.2196/48166

Freeman, A. L. J., Kerr, J., Recchia, G., Schneider, C. R., Lawrence, A. C. E., Finikarides, L., Luoni, G., Dryhurst, S., & Spiegelhalter, D. (2021). Communicating personalized risks from COVID-19: Guidelines from an empirical study. *Royal Society Open Science, 8*(4), 201721. https://doi.org/10.1098/rsos.201721

HLS19 Consortium of the WHO Action Network M-POHL. (2021). *International report on the methodology, results, and recommendations of the European Health Literacy Population Survey 2019–2021 (HLS19) of M-POHL*. Austrian National Public Health Institute. https://m-pohl.net/sites/m-pohl.net/files/inline-files/HLS19%20International%20Report.pdf

Kebede, A. S., Ozolins, L.-L., Holst, H., & Galvin, K. (2022). Digital engagement of older adults: Scoping review. *Journal of Medical Internet Research, 24*(12), e40192. https://doi.org/10.2196/40192

Kerr, J. R., Schneider, C. R., Freeman, A. L. J., Marteau, T., & van der Linden, S. (2022). Transparent communication of evidence does not undermine public trust in evidence. *PNAS Nexus, 1*(5), pgac280. https://doi.org/10.1093/pnasnexus/pgac280

Lee, K. J., Tilling, K. M., Cornish, R. P., Little, R. J. A., Bell, M. L., Goetghebeur, E., Hogan, J. W., & Carpenter, J. R. (2021). Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework. *Journal of Clinical Epidemiology, 134,* 79–88. https://doi.org/10.1016/j.jclinepi.2021.01.008

Lyles, C. R., Nguyen, O. K., Khoong, E. C., Aguilera, A., & Sarkar, U. (2023). Multilevel determinants of digital health equity: A literature synthesis to advance the field. *Annual Review of Public Health, 44,* 383–405. https://doi.org/10.1146/annurev-publhealth-071521-023913

Maini, I., Gilotra, K., & Sadigh, G. (2025). Patient factors associated with the use of online portal health information in the postpandemic era: Cross-sectional analysis of a national survey. *Journal of Medical Internet Research, 27,* e60472. https://doi.org/10.2196/60472

Mashhadi, S. F., Hisam, A., Sikander, S., Rathore, M. A., Rifaq, F., Khan, S. A., & Hafeez, A. (2021). Post discharge mHealth and teach-back communication effectiveness on hospital readmissions: A systematic review. *International Journal of Environmental Research and Public Health, 18*(19), 10442. https://doi.org/10.3390/ijerph181910442

Nan, X., Wang, Y., & Thier, K. (2022). Why do people believe health misinformation and who is at risk? A systematic review of individual differences in susceptibility to health misinformation. *Social Science & Medicine, 314,* 115398. https://doi.org/10.1016/j.socscimed.2022.115398

National Cancer Institute. (2023). *Health Information National Trends Survey 6 (HINTS 6): Methodology report*. https://hints.cancer.gov/docs/methodologyreports/HINTS_6_MethodologyReport.pdf

National Cancer Institute. (2025). *HINTS 6 public-use dataset and supporting documentation* [Data set; release updated August 2025]. https://hints.cancer.gov/data/download-data.aspx

National Center for Education Statistics. (2024, December 10). *U.S. adults score on par with international average in literacy skills, below international average in numeracy and adaptive problem solving* [Press release]. Institute of Education Sciences. https://ies.ed.gov/learn/press-release/u-s-adults-score-par-international-average-literacy-skills-below-international-average-numeracy-and

Nutbeam, D., & Lloyd, J. E. (2021). Understanding and responding to health literacy as a social determinant of health. *Annual Review of Public Health, 42*, 159–173. https://doi.org/10.1146/annurev-publhealth-090419-102529

Okan, O., Messer, M., Levin-Zamir, D., Paakkari, L., & Sørensen, K. (2023). Health literacy as a social vaccine in the COVID-19 pandemic. *Health Promotion International, 38*(4), daab197. https://doi.org/10.1093/heapro/daab197

Organisation for Economic Co-operation and Development. (2024). *Do adults have the skills they need to thrive in a changing world? Survey of Adult Skills 2023*. OECD Publishing. https://doi.org/10.1787/b263dc5d-en

Padilla, L., Hosseinpour, H., Fygenson, R., Howell, J., Chunara, R., & Bertini, E. (2022). Impact of COVID-19 forecast visualizations on pandemic risk perceptions. *Scientific Reports*, *12*, 2014. https://doi.org/10.1038/s41598-022-05353-1

Richter, G., & Krawczak, M. (2024). How to elucidate consent-free research use of medical data: A case for "health data literacy". *JMIR Medical Informatics*, *12*, e51350. https://doi.org/10.2196/51350

Richwine, C., Johnson, C., & Patel, V. (2023). Disparities in patient portal access and the role of providers in encouraging access and use. *Journal of the American Medical Informatics Association, 30*(2), 308–317. https://doi.org/10.1093/jamia/ocac227

Rodriguez, J. A., Saadi, A., Schwamm, L. H., Bates, D. W., & Samal, L. (2021). Disparities in telehealth use among California patients with limited English proficiency. *Health Affairs, 40*(3), 487–495. https://doi.org/10.1377/hlthaff.2020.00823

Santana, S., Brach, C., Harris, L., Ochiai, E., Blakey, C., Bevington, F., Kleinman, D., & Pronk, N. (2021). Updating health literacy for Healthy People 2030: Defining its importance for a new decade in public health. *Journal of Public Health Management and Practice, 27*(Suppl. 6), S258–S264. https://doi.org/10.1097/PHH.0000000000001324

Savitz, D. A., & Wellenius, G. A. (2023). Can cross-sectional studies contribute to causal inference? It depends. *American Journal of Epidemiology, 192*(4), 514–516. https://doi.org/10.1093/aje/kwac037

Schillinger, D. (2021). Social determinants, health literacy, and disparities: Intersections and controversies. *HLRP: Health Literacy Research and Practice, 5*(3), e234–e243. https://doi.org/10.3928/24748307-20210712-01

Schneider, C. R., Freeman, A. L. J., Spiegelhalter, D., & van der Linden, S. (2021). The effects of quality of evidence communication on perception of public health information about COVID-19: Two randomised controlled trials. *PLOS ONE, 16*(11), e0259048. https://doi.org/10.1371/journal.pone.0259048

Shahid, R., Shoker, M., Chu, L. M., Frehlick, R., Ward, H., & Pahwa, P. (2022). Impact of low health literacy on patients' health outcomes: A multicenter cohort study. *BMC Health Services Research, 22*, 1148. https://doi.org/10.1186/s12913-022-08527-9

Sieck, C. J., Sheon, A., Ancker, J. S., Castek, J., Callahan, B., & Siefer, A. (2021). Digital inclusion as a social determinant of health. *npj Digital Medicine, 4*, 52. https://doi.org/10.1038/s41746-021-00413-8

Sørensen, K., Levin-Zamir, D., Duong, T. V., Okan, O., Brasil, V. V., & Nutbeam, D. (2021). Building health literacy system capacity: A framework for health literate systems. *Health Promotion International, 36*(Suppl. 1), i13–i23. https://doi.org/10.1093/heapro/daab153

Steitz, B. D., Sulieman, L., Wright, A., & Rosenbloom, S. T. (2021). Association of immediate release of test results to patients with implications for clinical workflow. *JAMA Network Open, 4*(10), e2129553. https://doi.org/10.1001/jamanetworkopen.2021.29553

Steitz, B. D., Turer, R. W., Lin, C.-T., MacDonald, S., Salmi, L., Wright, A., Lehmann, C. U., Langford, K., McDonald, S. A., Reese, T. J., Sternberg, P., Chen, Q., Rosenbloom, S. T., & DesRoches, C. M. (2023). Perspectives of patients about immediate access to test results through an online patient portal. *JAMA Network Open, 6*(3), e233572. https://doi.org/10.1001/jamanetworkopen.2023.3572

van Kessel, R., Wong, B. L. H., Clemens, T., & Brand, H. (2022). Digital health literacy as a super determinant of health: More than simply the sum of its parts. *Internet Interventions, 27*, 100500. https://doi.org/10.1016/j.invent.2022.100500

Wang, X., Shi, J., & Kong, H. (2021). Online health information seeking: A review and meta-analysis. *Health Communication, 36*(10), 1163–1175. https://doi.org/10.1080/10410236.2020.1748829

World Health Organization. (2022). *Health literacy development for the prevention and control of noncommunicable diseases: Volume 1. Overview*. https://www.who.int/publications/i/9789240055339

Yao, R., Zhang, W., Evans, R., Cao, G., Rui, T., & Shen, L. (2022). Inequities in health care services caused by the adoption of digital health technologies: Scoping review. *Journal of Medical Internet Research, 24*(3), e34144. https://doi.org/10.2196/34144

Zhong, H., Song, Y., Liu, Q., & Chen, H. (2025). Factors influencing eHealth literacy worldwide: Systematic review and meta-analysis. *Journal of Medical Internet Research, 27*, e50313. https://doi.org/10.2196/50313

Appendix A. Analysis Code

This appendix reproduces the analysis script used for every estimate, table and figure reported in this article. The screenshots show the actual source file (hints6_hdl_analysis.py) with original line numbers; long lines are soft-wrapped for display. The script reads the official HINTS 6 public-use Stata file (release updated August 2025) and applies the final person weight (PERSON_FINWT0) with 50 jackknife replicate weights (PERSON_FINWT1–50, multiplier 0.98, 49 degrees of freedom). The original descriptive and primary analyses were executed in Python 3.12.14 (execution log supplied with the analysis package). The extended analyses shown here were run in Python 3.13.16 and re-run in Python 3.12.13 (NumPy 2.5.4, pandas 3.0.6, SciPy 1.18.1, Matplotlib 3.11.2), with identical results. Lines 355–474, which draw Figures 1–6, are omitted for space and are included in the supplementary code file.

hints6_hdl_analysis.py — lines 1–46: study header, data import and survey design

```
"""
Health Data Literacy as a Public Health Intervention
Complete, reproducible analysis of the HINTS 6 (2022) public-use file.

Run:   python hints6_hdl_analysis.py
Needs: numpy, pandas, scipy, matplotlib
Data:  source/HINTS6_STATA_20250731/hints6_public.dta  (NCI HINTS 6 public-use release, updated Aug 2025
    )

Survey design: full-sample weight PERSON_FINWT0 for point estimates; 50 jackknife replicate
weights (PERSON_FINWT1-50), JK1 multiplier 0.98, 49 degrees of freedom, as recommended by NCI.

Analyses
  A. Descriptive prevalence (overall, by subgroup, full item distributions)
  B. Main survey-weighted logistic model of difficulty (demographics + Internet use)
     + design-based joint Wald F tests for each multi-category factor
  C. Extended model adding health status, insurance, rurality and survey language
  D. Education x Internet-use interaction model and adjusted predicted probabilities
  E. Survey-weighted proportional-odds (ordinal) model of the four-level item
  F. Clinical-communication models (clinicians explained clearly / helped with uncertainty)
  G. Secondary information outcomes: Model A (demographics) and Model B (+ health and care use),
     Benjamini-Hochberg q-values, and dose-response gradient across the four response levels
  H. Sensitivity: omit income, very-hard-only outcome, complete vs excluded cases, missing-outcome bound
    s
  J. Proportional-odds (parallel-lines) assumption check for the ordinal model
  I. Figures 1-6
"""
from pathlib import Path
import json, warnings
import numpy as np
import pandas as pd
from scipy.special import expit
from scipy.stats import t, f as fdist
from scipy.optimize import minimize

warnings.filterwarnings("ignore")
ROOT = Path(__file__).resolve().parent
OUT = ROOT / "results"; OUT.mkdir(exist_ok=True)
FIG = ROOT / "figures"; FIG.mkdir(exist_ok=True)
SOURCE = ROOT / "source/HINTS6_STATA_20250731/hints6_public.dta"

raw = pd.read_stata(SOURCE, convert_categoricals=False)
assert len(raw) == 6252
WT = [f"PERSON_FINWT{i}" for i in range(51)]
W = raw[WT].to_numpy(float)
DF, SCALE = 49, 0.98
CRIT = t.ppf(0.975, DF)

```

Figure A1. Study header, data import and survey design settings (lines 1–46).

hints6_hdl_analysis.py — lines 47–101: variable coding and eligibility rules

```python
# ------------------------------------------------------------------------------ variables
def clean(v, valid):
    return raw[v].where(raw[v].isin(list(valid)))

def binary(v, valid, yes):
    s = clean(v, valid)
    return s.isin(yes).astype(float).where(s.notna())

d = pd.DataFrame(index=raw.index)
d["medical_stats"] = clean("UndMedicalStats", range(1, 5))          # 1 very easy ... 4 very hard
d["difficulty"] = binary("UndMedicalStats", range(1, 5), [3, 4])
d["very_hard"] = binary("UndMedicalStats", range(1, 5), [4])
d["age"] = pd.cut(raw.Age.where(raw.Age.between(18, 120)), [17, 34, 49, 64, 120], labels=[1, 2, 3, 4]).astype(float)
d["education"] = clean("EducA", range(1, 5))
d["income"] = clean("HHInc", range(1, 6))
d["sex"] = clean("BirthGender", range(1, 3))
d["race"] = clean("RaceEthn5", range(1, 6))
d["internet"] = binary("UseInternet", [1, 2], [1])
# extended covariates
gh = clean("GeneralHealth", range(1, 6))
d["health"] = gh.map({1: 1, 2: 1, 3: 2, 4: 3, 5: 3})                  # 1 excellent/very good, 2 good, 3 fair/poor
d["insured"] = binary("HealthInsurance2", [1, 2], [1])
d["rural"] = raw.RUC2013.where(raw.RUC2013.between(1, 9)).map(lambda x: 1.0 if x >= 4 else 0.0)
d["spanish"] = raw.LANGUAGE_FLAG.map({1: 0.0, 2: 1.0})
fv = clean("FreqGoProvider", range(0, 7))
d["visits"] = fv.map({0: 1, 1: 2, 2: 2, 3: 3, 4: 3, 5: 4, 6: 4})      # 0, 1-2, 3-4, 5+
d["explain_not_always"] = binary("ExplainedClearly", range(1, 5), [2, 3, 4])
d["uncert_not_always"] = binary("HelpUncertainty", range(1, 5), [2, 3, 4])
# secondary outcomes (eligibility from official routing)
d["forms_low"] = binary("ConfidentMedForms", range(1, 5), [3, 4])
d["web_low"] = binary("ConfidentInternetHealth", range(1, 6), [4, 5])
d["cancer_hard"] = binary("CancerTooHardUnderstand", range(1, 5), [1, 2]).where(raw.SeekCancerInfo.eq(1))
d["online_seek"] = binary("Electronic2_HealthInfo", [1, 2], [1]).where(raw.UseInternet.eq(1))
d["portal_hard"] = binary("UnderstandOnlineMedRec2", range(1, 5), [3, 4]).where(raw.AccessOnlineRecord2.isin([1, 2, 3, 4]))

CATS = {
    "age": {1: "18-34 years", 2: "35-49 years", 3: "50-64 years", 4: "65 years or older"},
    "education": {1: "Less than high school", 2: "High school graduate", 3: "Some college", 4: "College graduate or more"},
    "income": {1: "Less than $20,000", 2: "$20,000-$34,999", 3: "$35,000-$49,999", 4: "$50,000-$74,999", 5: "$75,000 or more"},
    "sex": {1: "Male", 2: "Female"},
    "race": {1: "Non-Hispanic White", 2: "Non-Hispanic Black", 3: "Hispanic", 4: "Non-Hispanic Asian", 5: "Non-Hispanic Other"},
    "internet": {0: "Does not use the Internet", 1: "Uses the Internet"},
    "health": {1: "Excellent/very good", 2: "Good", 3: "Fair/poor"},
    "insured": {0: "Uninsured", 1: "Insured"},
    "rural": {0: "Metropolitan", 1: "Non-metropolitan"},
    "spanish": {0: "English", 1: "Spanish"},
    "visits": {1: "No visits", 2: "1-2 visits", 3: "3-4 visits", 4: "5 or more visits"},
    "difficulty": {0: "Easy", 1: "Hard/very hard"},
    "explain_not_always": {0: "Always", 1: "Not always"},
    "uncert_not_always": {0: "Always", 1: "Not always"},
}
REF = {"age": 1, "education": 4, "income": 5, "sex": 1, "race": 1, "internet": 1, "health": 1, "insured": 1,
       "rural": 0, "spanish": 0, "visits": 1, "difficulty": 0, "explain_not_always": 0, "uncert_not_always": 0}
COVS = ["age", "education", "income", "sex", "race", "internet"]

```

Figure A2. Variable coding, HINTS eligibility (routing) rules and reference categories (lines 47–101).

hints6_hdl_analysis.py — lines 102–151: weighted prevalence, design matrix, IRLS

```
# ------------------------------------------------------------------------------ estimators
def proportion(series, domain=None):
    m = series.notna().to_numpy().copy()
    if domain is not None:
        m &= np.asarray(domain, dtype=bool)
    y = series.to_numpy(float)[m]; w = W[m]
    est = (w * y[:, None]).sum(0) / w.sum(0)
    se = np.sqrt(SCALE * np.sum((est[1:] - est[0]) ** 2))
    return {"n": int(m.sum()), "events": int(y.sum()), "pct": 100 * est[0], "se": 100 * se,
            "lo": 100 * max(0, est[0] - CRIT * se), "hi": 100 * min(1, est[0] + CRIT * se),
            "rse": 100 * se / est[0] if est[0] > 0 else np.nan}

def design(terms, mask, frame=None, check=True):
    """terms: names of categorical covariates, ('num', name) for a numeric term,
    or ('int', a, b) for an interaction between two categorical covariates."""
    frame = d if frame is None else frame
    cols, names = [np.ones(mask.sum())], ["Intercept"]
    dummies = {}
    for term in terms:
        if isinstance(term, tuple) and term[0] == "num":
            cols.append(frame.loc[mask, term[1]].to_numpy(float)); names.append(term[1]); continue
        if isinstance(term, tuple) and term[0] == "int":
            _, a, b = term
            for na, va in dummies[a]:
                for nb, vb in dummies[b]:
                    cols.append(va * vb); names.append(f"{na}*{nb}")
            continue
        dummies[term] = []
        for lev in CATS[term]:
            if lev != REF[term]:
                v = (frame.loc[mask, term].to_numpy() == lev).astype(float)
                cols.append(v); names.append(f"{term}={lev}"); dummies[term].append((f"{term}={lev}", v)
    )
    X = np.column_stack(cols)
    if check: assert np.linalg.matrix_rank(X) == X.shape[1], "rank deficient"
    return X, names

def irls(X, y, w):
    w = w / w.mean(); b = np.zeros(X.shape[1])
    for _ in range(100):
        eta = X @ b; p = expit(eta)
        step = np.linalg.solve(X.T @ ((w * p * (1 - p))[:, None] * X), X.T @ (w * (y - p)))
        loss = np.sum(w * (np.logaddexp(0, eta) - y * eta)); fac = 1.0
        while fac > 1e-8:
            c = b + fac * step; ce = X @ c
            if np.sum(w * (np.logaddexp(0, ce) - y * ce)) <= loss + 1e-9: break
            fac *= 0.5
        b = c
        if np.max(np.abs(fac * step)) < 1e-10: break
    return b

```

Figure A3. Survey-weighted prevalence with jackknife standard errors, design-matrix construction and the weighted logistic (IRLS) estimator (lines 102–151).

hints6_hdl_analysis.py — lines 152–190: jackknife logistic model, Wald test, margins

```python
class Fit:
    pass

def logit(outcome, terms, domain=None, name=""):
    needed = [outcome] + [x for tm in terms for x in ([tm] if isinstance(tm, str) else tm[1:])]
    mask = d[list(dict.fromkeys(needed))].notna().all(axis=1).to_numpy().copy()
    if domain is not None: mask &= np.asarray(domain, dtype=bool)
    X, names = design(terms, mask); y = d.loc[mask, outcome].to_numpy(float); w = W[mask]
    b = irls(X, y, w[:, 0]); brep = np.array([irls(X, y, w[:, r]) for r in range(1, 51)])
    cov = SCALE * (brep - b).T @ (brep - b); se = np.sqrt(np.diag(cov))
    fit = Fit(); fit.__dict__.update(dict(name=name, outcome=outcome, terms=terms, mask=mask, X=X, y=y,
    w=w,
                                          b=b, brep=brep, cov=cov, se=se, names=names))
    fit.table = pd.DataFrame({"model": name, "outcome": outcome, "term": names, "n": int(mask.sum()),
                              "events": int(y.sum()), "coef": b, "se": se, "OR": np.exp(b),
                              "lo": np.exp(b - CRIT * se), "hi": np.exp(b + CRIT * se),
                              "p": 2 * t.sf(np.abs(b / se), DF)})
    return fit

def wald(fit, prefix):
    """Design-based joint Wald test, adjusted F with (k, df-k+1) degrees of freedom."""
    idx = [i for i, n in enumerate(fit.names) if n.startswith(prefix)]
    k = len(idx); bb = fit.b[idx]; V = fit.cov[np.ix_(idx, idx)]
    W_ = float(bb @ np.linalg.solve(V, bb)); F = W_ / k * (DF - k + 1) / DF
    return {"term": prefix.rstrip("="), "k": k, "F": F, "df1": k, "df2": DF - k + 1, "p": float(fdist.sf
    (F, k, DF - k + 1))}

def margins(fit, var, levels, frame_mod=None):
    """Average adjusted predicted probability (marginal standardisation) for each level of var."""
    rows = []
    base = d.loc[fit.mask].copy()
    for lev in levels:
        f2 = d.copy(); f2.loc[fit.mask, var] = lev
        if frame_mod: f2 = frame_mod(f2, lev)
        X2, _ = design(fit.terms, fit.mask, frame=f2, check=False)
        est = [np.average(expit(X2 @ fit.b), weights=fit.w[:, 0])]
        est += [np.average(expit(X2 @ fit.brep[r - 1]), weights=fit.w[:, r]) for r in range(1, 51)]
        est = np.array(est); se = np.sqrt(SCALE * np.sum((est[1:] - est[0]) ** 2))
        rows.append({"variable": var, "level": lev, "pct": 100 * est[0], "lo": 100 * (est[0] - CRIT * se
    ), "hi": 100 * (est[0] + CRIT * se)})
    return rows

```

Figure A4. Jackknife-replicate logistic model, design-based joint Wald F test and adjusted predicted probabilities by marginal standardisation (lines 152–190).

hints6_hdl_analysis.py — lines 191–251: descriptives, main, extended and interaction models

```python
# ---------------------------------------------------------------------------- A. descriptives
valid = d.difficulty.notna()
res = {"n_total": len(raw), "n_valid": int(valid.sum()), "n_missing_outcome": int((~valid).sum())}
res["overall"] = proportion(d.difficulty); res["very_hard"] = proportion(d.very_hard)

t1 = []
for c in COVS + ["health", "insured", "rural", "spanish"]:
    for k, lab in CATS[c].items():
        comp = proportion((d[c] == k).astype(float).where(d[c].notna()), valid)
        prev = proportion(d.difficulty, d[c].eq(k))
        t1.append({"variable": c, "category": lab, "n": int((valid & d[c].eq(k)).sum()), "sample_pct": comp["pct"],
                   "diff_n": prev["events"], "diff_pct": prev["pct"], "diff_lo": prev["lo"], "diff_hi": prev["hi"],
                   "rse": prev["rse"], "missing_in_valid": int((valid & d[c].isna()).sum())})
t1 = pd.DataFrame(t1); t1.to_csv(OUT / "T2_sample_and_prevalence.csv", index=False)

dist = []
items = [("medical_stats", "UndMedicalStats", range(1, 5), None, ["Very easy", "Easy", "Hard", "Very hard"]),
         ("forms", "ConfidentMedForms", range(1, 5), None, ["Very", "Somewhat", "A little", "Not at all"]),
         ("web", "ConfidentInternetHealth", range(1, 6), None, ["Completely", "Very", "Somewhat", "A little", "Not at all"]),
         ("cancer", "CancerTooHardUnderstand", range(1, 5), raw.SeekCancerInfo.eq(1), ["Strongly agree", "Somewhat agree", "Somewhat disagree", "Strongly disagree"]),
         ("portal", "UnderstandOnlineMedRec2", range(1, 5), raw.AccessOnlineRecord2.isin([1, 2, 3, 4]), ["Very easy", "Somewhat easy", "Somewhat difficult", "Very difficult"])]
for key, v, levs, dom, labs in items:
    s = clean(v, levs)
    for lev, lab in zip(levs, labs):
        r = proportion(s.eq(lev).astype(float).where(s.notna()), dom)
        dist.append({"item": key, "response": lab, "n": r["events"], "denominator": r["n"], "pct": r["pct"], "lo": r["lo"], "hi": r["hi"]})
dist = pd.DataFrame(dist); dist.to_csv(OUT / "item_distributions.csv", index=False)

# ---------------------------------------------------------------------------- B. main model
main = logit("difficulty", COVS, name="Main model")
main_wald = [wald(main, p) for p in ["age=", "education=", "income=", "sex=", "race=", "internet="]]
# linear trend across education (score 1-4) in the same adjusted model
d["edu_score"] = d.education
trend = logit("difficulty", ["age", ("num", "edu_score"), "income", "sex", "race", "internet"], name="Education trend")
# OR per one step DOWN the education ladder = 1/OR for score increasing
r_ = trend.table.loc[trend.table.term == "edu_score"].iloc[0]
res["edu_trend_OR_per_level_down"] = {"OR": 1 / r_.OR, "lo": 1 / r_.hi, "hi": 1 / r_.lo, "p": r_.p}

# ---------------------------------------------------------------------------- C. extended model
ext = logit("difficulty", COVS + ["health", "insured", "rural", "spanish"], name="Extended model")
ext_wald = [wald(ext, p) for p in ["education=", "income=", "race=", "health=", "internet="]]

# ---------------------------------------------------------------------------- D. interaction + margins
inter = logit("difficulty", COVS + [("int", "education", "internet")], name="Education x Internet")
idx = [i for i, n in enumerate(inter.names) if "*" in n]
bb = inter.b[idx]; V = inter.cov[np.ix_(idx, idx)]; k = len(idx)
Fi = float(bb @ np.linalg.solve(V, bb)) / k * (DF - k + 1) / DF
res["interaction_test"] = {"F": Fi, "df1": k, "df2": DF - k + 1, "p": float(fdist.sf(Fi, k, DF - k + 1)), "n": int(inter.mask.sum())}
marg_main = margins(main, "education", [1, 2, 3, 4]) + margins(main, "internet", [0, 1])
marg_cross = []
for net in [0, 1]:
    for e in [1, 2, 3, 4]:
        f2 = d.copy(); f2.loc[inter.mask, "education"] = e; f2.loc[inter.mask, "internet"] = net
        X2, _ = design(inter.terms, inter.mask, frame=f2, check=False)
        est = np.array([np.average(expit(X2 @ inter.b), weights=inter.w[:, 0])] +
                       [np.average(expit(X2 @ inter.brep[r - 1]), weights=inter.w[:, r]) for r in range(1, 51)])
        se = np.sqrt(SCALE * np.sum((est[1:] - est[0]) ** 2))
        marg_cross.append({"internet": net, "education": e, "pct": 100 * est[0], "lo": 100 * (est[0] - CRIT * se), "hi": 100 * (est[0] + CRIT * se)})
pd.DataFrame(marg_main).to_csv(OUT / "margins_main.csv", index=False)
pd.DataFrame(marg_cross).to_csv(OUT / "margins_education_by_internet.csv", index=False)

```

Figure A5. Descriptive estimates, main model, education trend, extended model and education × Internet interaction with predicted probabilities (lines 191–251).

hints6_hdl_analysis.py — lines 252–287: proportional-odds and clinical-communication models

```
# ---------------------------------------------------------------------------- E. ordinal model
def ordinal_fit(X, y, w):
    """Proportional-odds logit: P(Y<=j)=expit(theta_j - Xb); Y in 1..4. Positive b = harder."""
    X = X[:, 1:]; w = w / w.mean(); J = 4
    def unpack(z):
        th = np.cumsum(np.concatenate([[z[0]], np.exp(z[1:J - 1])])); return th, z[J - 1:]
    def nll(z):
        th, b = unpack(z); eta = X @ b
        lo = np.where(y == 1, 0.0, expit(th[np.clip(y - 2, 0, J - 2).astype(int)] - eta))
        up = np.where(y == J, 1.0, expit(th[np.clip(y - 1, 0, J - 2).astype(int)] - eta))
        return -np.sum(w * np.log(np.clip(up - lo, 1e-300, None)))
    z0 = np.concatenate([[-1.5, 0.9, 0.9], np.zeros(X.shape[1])])
    r = minimize(nll, z0, method="BFGS", options={"gtol": 1e-6, "maxiter": 2000})
    return r.x[J - 1:]

om = d[["medical_stats"] + COVS].notna().all(axis=1).to_numpy()
Xo, onames = design(COVS, om); yo = d.loc[om, "medical_stats"].to_numpy(float); wo = W[om]
bo = ordinal_fit(Xo, yo, wo[:, 0]); bor = np.array([ordinal_fit(Xo, yo, wo[:, r]) for r in range(1, 51)])
seo = np.sqrt(np.diag(SCALE * (bor - bo).T @ (bor - bo)))
ordinal = pd.DataFrame({"term": onames[1:], "OR": np.exp(bo), "lo": np.exp(bo - CRIT * seo), "hi": np.exp(bo + CRIT * seo),
                        "p": 2 * t.sf(np.abs(bo / seo), DF), "n": int(om.sum())})
ordinal.to_csv(OUT / "ordinal_model.csv", index=False)

# ---------------------------------------------------------------------------- F. clinical communication
comm = []
for exp_var in ["explain_not_always", "uncert_not_always"]:
    dom = d.visits.isin([2, 3, 4]).to_numpy()
    f1 = logit("difficulty", [exp_var] + COVS + ["health"], domain=dom, name=f"Communication {exp_var}")
    row = f1.table[f1.table.term == f"{exp_var}=1"].iloc[0].to_dict()
    for lev in [0, 1]:
        pr = proportion(d.difficulty, d[exp_var].eq(lev) & d.visits.isin([2, 3, 4]))
        row[f"pct_{lev}"] = pr["pct"]; row[f"n_{lev}"] = pr["n"]; row[f"lo_{lev}"] = pr["lo"]; row[f"hi_{lev}"] = pr["hi"]
    row["exposure_prev"] = proportion(d[exp_var], dom)["pct"]
    comm.append(row)
comm = pd.DataFrame(comm); comm.to_csv(OUT / "communication_models.csv", index=False)

```

Figure A6. Survey-weighted proportional-odds (ordinal) model and clinician-communication models (lines 252–287).

hints6_hdl_analysis.py — lines 288–330: information outcomes, gradient, FDR and sensitivity

```
# ------------------------------------------------------------------------------ G. secondary outcomes
OUTS = ["forms_low", "web_low", "cancer_hard", "online_seek", "portal_hard"]
sec = []; full_sec = []
for o in OUTS:
    base = [c for c in COVS if not (o in ["online_seek", "portal_hard"] and c == "internet")]
    fa = logit(o, ["difficulty"] + base, name=f"{o} model A")
    fb = logit(o, ["difficulty"] + base + ["health", "insured", "rural", "visits"], name=f"{o} model B")
    fg = logit(o, [("num", "medical_stats")] + base, name=f"{o} gradient")
    full_sec += [fa.table, fb.table, fg.table]
    ra = fa.table[fa.table.term == "difficulty=1"].iloc[0]; rb = fb.table[fb.table.term == "difficulty=1
    "].iloc[0]
    rg = fg.table[fg.table.term == "medical_stats"].iloc[0]
    row = {"outcome": o, "A_n": ra.n, "A_OR": ra.OR, "A_lo": ra.lo, "A_hi": ra.hi, "A_p": ra.p,
           "B_n": rb.n, "B_OR": rb.OR, "B_lo": rb.lo, "B_hi": rb.hi, "B_p": rb.p,
           "G_OR": rg.OR, "G_lo": rg.lo, "G_hi": rg.hi, "G_p": rg.p}
    for k_, lab in [(0, "easy"), (1, "hard")]:
        pr = proportion(d[o], d.difficulty.eq(k_)); row[f"{lab}_pct"] = pr["pct"]; row[f"{lab}_n"] = pr[
    "n"]
    for lev in [1, 2, 3, 4]:
        pr = proportion(d[o], d.medical_stats.eq(lev))
        row[f"L{lev}_pct"] = pr["pct"]; row[f"L{lev}_lo"] = pr["lo"]; row[f"L{lev}_hi"] = pr["hi"]; row[
    f"L{lev}_n"] = pr["n"]
    sec.append(row)
sec = pd.DataFrame(sec)
def bh(p):
    p = np.asarray(p); o = np.argsort(p); r = p[o] * len(p) / np.arange(1, len(p) + 1)
    q = np.minimum.accumulate(r[::-1])[::-1]; out = np.empty(len(p)); out[o] = np.minimum(1, q); return
    out
sec["A_q"] = bh(sec.A_p); sec["B_q"] = bh(sec.B_p)
sec.to_csv(OUT / "secondary_outcomes.csv", index=False)
pd.concat(full_sec).to_csv(OUT / "secondary_full_models.csv", index=False)

# ------------------------------------------------------------------------------ H. sensitivity
noinc = logit("difficulty", [c for c in COVS if c != "income"], name="Omit income")
vh = logit("very_hard", COVS, name="Very hard only")
cc = proportion(d.difficulty, main.mask); ex = proportion(d.difficulty, valid.to_numpy() & ~main.mask)
mw = W[~valid.to_numpy(), 0].sum() / W[:, 0].sum(); ov = res["overall"]["pct"]
res.update({"complete_case": cc, "excluded_case": ex, "missing_outcome_weight_pct": 100 * mw,
            "bounds": [ov * (1 - mw), ov * (1 - mw) + 100 * mw],
            "main_n": int(main.mask.sum()), "main_events": int(main.y.sum()),
            "ext_n": int(ext.mask.sum()), "ext_events": int(ext.y.sum()),
            "noinc_n": int(noinc.mask.sum()), "vh_n": int(vh.mask.sum()), "vh_events": int(vh.y.sum()),
            "ordinal_n": int(om.sum()),
            "main_wald": main_wald, "ext_wald": ext_wald})
pd.concat([main.table, trend.table, ext.table, inter.table, noinc.table, vh.table]).to_csv(OUT / "diffic
    ulty_models.csv", index=False)
(OUT / "summary.json").write_text(json.dumps(res, indent=2, default=float))

```

Figure A7. Information-outcome models (Model A, Model B and dose–response gradient), Benjamini–Hochberg q-values and sensitivity analyses (lines 288–330).

hints6_hdl_analysis.py — lines 331–354: proportional-odds (parallel-lines) assumption check

```
331 # ------------------------------------------------------------------------------ J. proportional-odds che
    ck
332 # Brant-type, design-based assessment: fit separate survey-weighted logits for each cumulative split
333 # (very easy vs rest; very easy/easy vs hard/very hard; very hard vs rest) on the same complete cases,
334 # then test equality of the slope vectors across splits with a jackknife Wald F test.
335 cuts = {}
336 for j in [1, 2, 3]:
337     yj = (yo > j).astype(float)
338     bj = irls(Xo, yj, wo[:, 0]); bjr = np.array([irls(Xo, yj, wo[:, r]) for r in range(1, 51)])
339     cuts[j] = (bj[1:], bjr[:, 1:], int(yj.sum()))
340 def po_test(idx):
341     d0 = np.concatenate([cuts[2][0][idx] - cuts[1][0][idx], cuts[3][0][idx] - cuts[1][0][idx]])
342     dr = np.hstack([cuts[2][1][:, idx] - cuts[1][1][:, idx], cuts[3][1][:, idx] - cuts[1][1][:, idx]])
343     V = SCALE * (dr - d0).T @ (dr - d0); k = len(d0)
344     Wst = float(d0 @ np.linalg.pinv(V) @ d0); F = Wst / k * (DF - k + 1) / DF
345     return {"k": k, "F": F, "df1": k, "df2": DF - k + 1, "p": float(fdist.sf(F, k, DF - k + 1))}
346 slope_names = onames[1:]
347 po_rows = [{"test": "Global (all covariates)", **po_test(list(range(len(slope_names))))}]
348 for v in COVS:
349     po_rows.append({"test": v, **po_test([i for i, n in enumerate(slope_names) if n.startswith(v + "=")]
    )})
350 po = pd.DataFrame(po_rows); po.to_csv(OUT / "proportional_odds_check.csv", index=False)
351 cut_or = pd.DataFrame({"term": slope_names, **{f"OR_cut{j}": np.exp(cuts[j][0]) for j in cuts}})
352 cut_or.to_csv(OUT / "proportional_odds_cutpoint_ORs.csv", index=False)
353 res["po_events_per_cut"] = {j: cuts[j][2] for j in cuts}
354
```

Figure A8. Design-based Brant-type test of the proportional-odds (parallel-lines) assumption across the three cumulative splits (lines 331–354).

Terminal — Python 3.12.13: python hints6_hdl_analysis.py (output 1 of 4)

```
$ python --version
Python 3.12.13
$ python hints6_hdl_analysis.py
{
 "overall": {
  "n": 6011,
  "events": 1589,
  "pct": 26.75706232998883,
  "se": 1.0318270234142957,
  "lo": 24.68352829473469,
  "hi": 28.830596365242968,
  "rse": 3.8562791785174513
 },
 "very_hard": {
  "n": 6011,
  "events": 174,
  "pct": 2.5564936873244877,
  "se": 0.35482725027582396,
  "lo": 1.843441631711533,
  "hi": 3.269545742937442,
  "rse": 13.879449498941277
 },
 "main_n": 5214,
 "main_events": 1331,
 "ext_n": 5163,
 "ext_events": 1315,
 "interaction_test": {
  "F": 0.7474995413652356,
  "df1": 3,
  "df2": 47,
  "p": 0.5292640496763001,
  "n": 5214
 },
 "edu_trend_OR_per_level_down": {
  "OR": 1.4192474892563742,
  "lo": 1.2105065875263359,
  "hi": 1.6639838696596099,
  "p": 5.415832022639227e-05
 },
 "bounds": [
  25.603725149612313,
  29.914128011304186
 ]
}
        term  k         F  df1  df2         p
0        age  3  1.474509    3   47  0.233570
1  education  3  8.659550    3   47  0.000111
2     income  4  1.445959    4   46  0.234153
3        sex  1  0.207301    1   49  0.650902
4       race  4  1.646873    4   46  0.178687
5   internet  1  9.920914    1   49  0.002784
        term  k          F  df1  df2             p
0  education  3   7.520488    3   47  3.283877e-04
1     income  4   1.399333    4   46  2.491485e-01
2       race  4   0.918317    4   46  4.614265e-01
3     health  2  19.721097    2   48  5.603825e-07
4   internet  1   7.633567    1   49  8.046247e-03
```

Figure B1. Console output (part 1): Python version, overall and very-hard prevalence, model sample sizes, interaction test, education trend, missing-data bounds and joint Wald tests for the main and extended models.

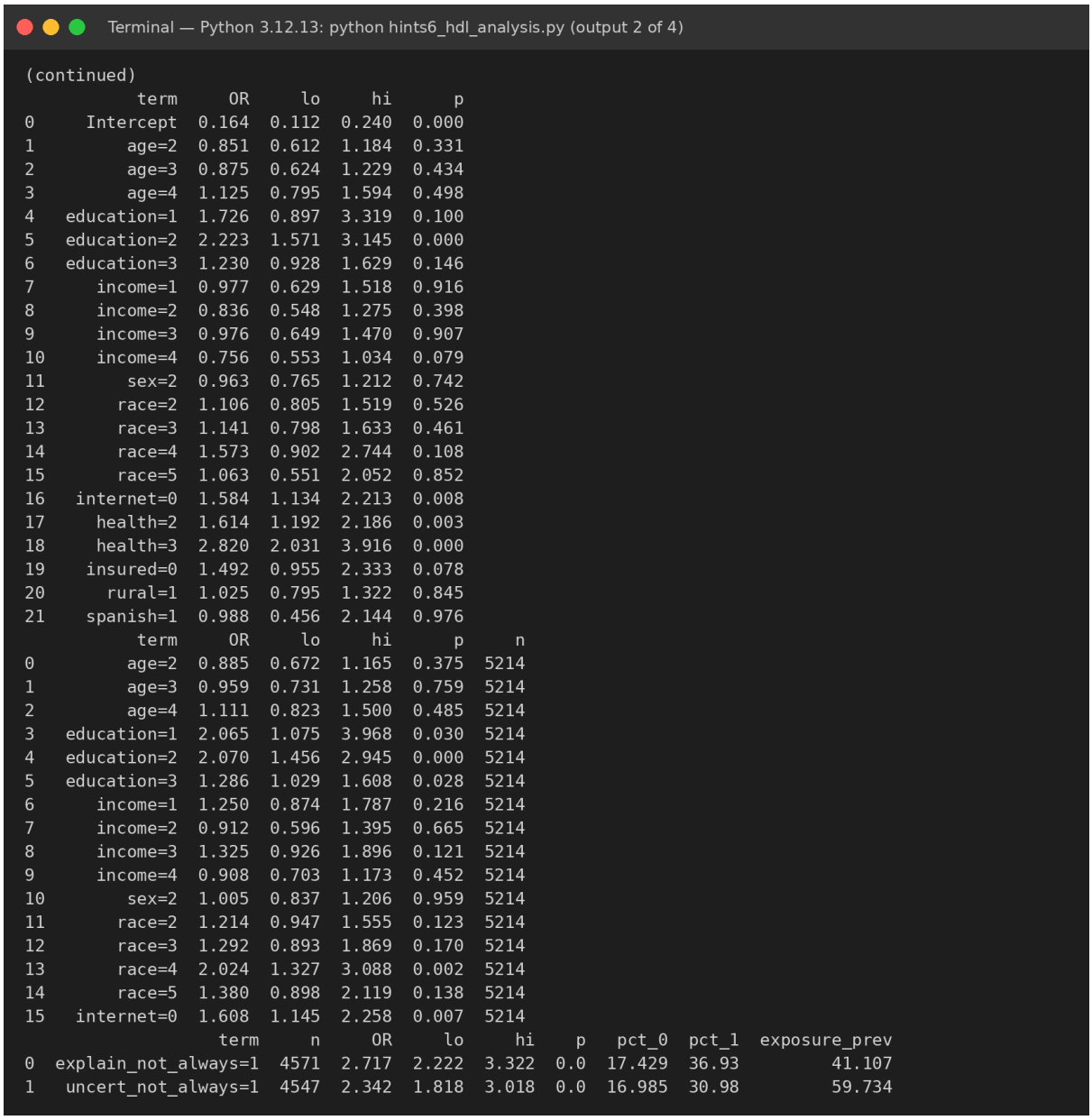


Figure B2. Console output (part 2): extended-model odds ratios, proportional-odds model and clinician-communication models.

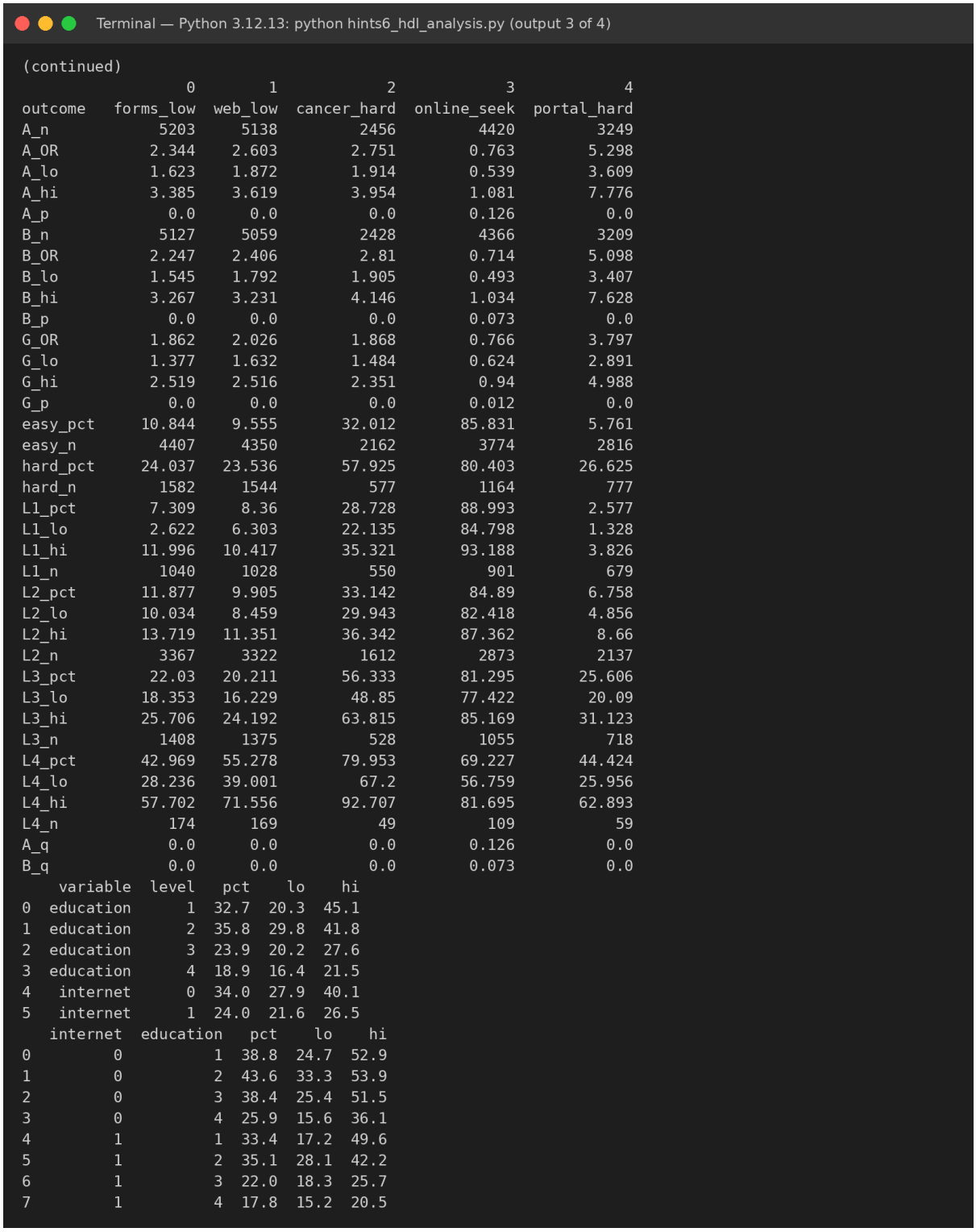


Figure B3. Console output (part 3): information-outcome models (A, B and gradient), outcome prevalence by response level, q-values and adjusted predicted probabilities.

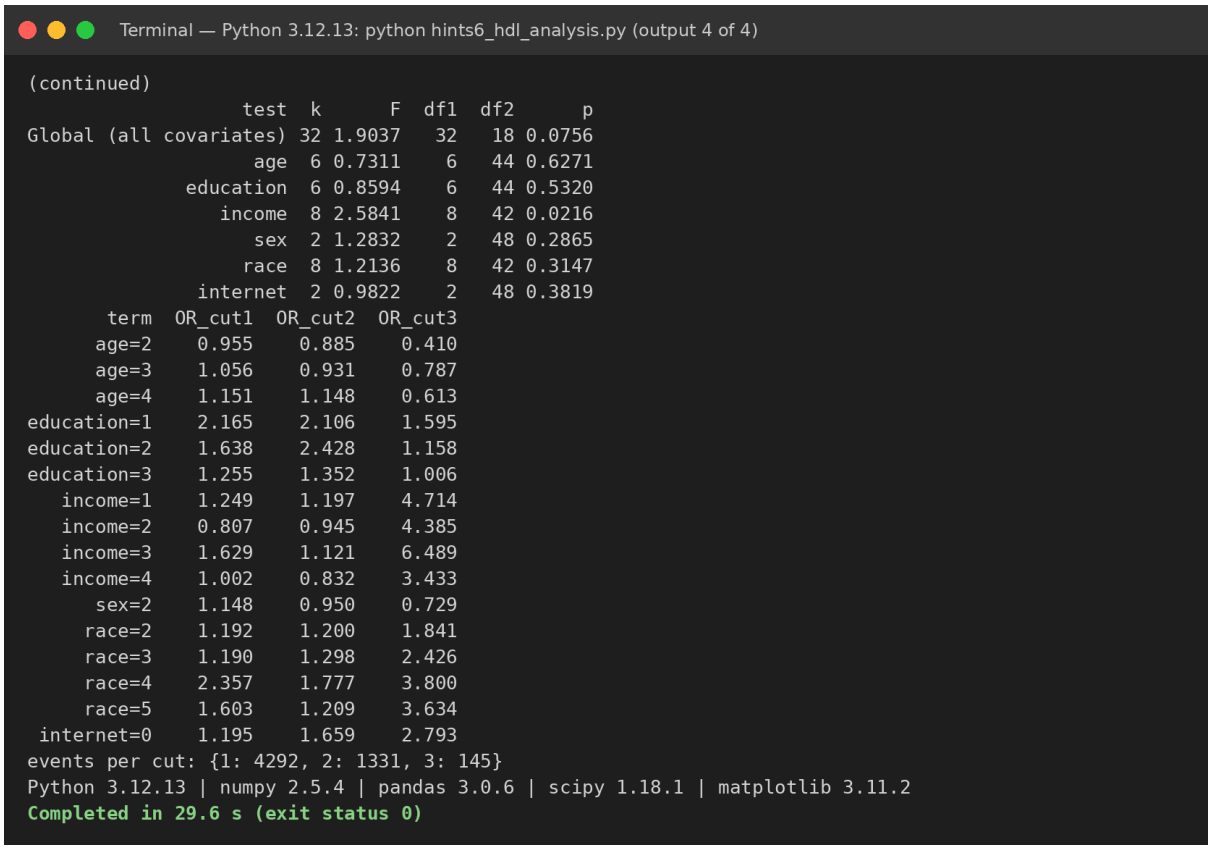


Figure B4. Console output (part 4): proportional-odds assumption tests, cutpoint-specific odds ratios, events per split and software versions.

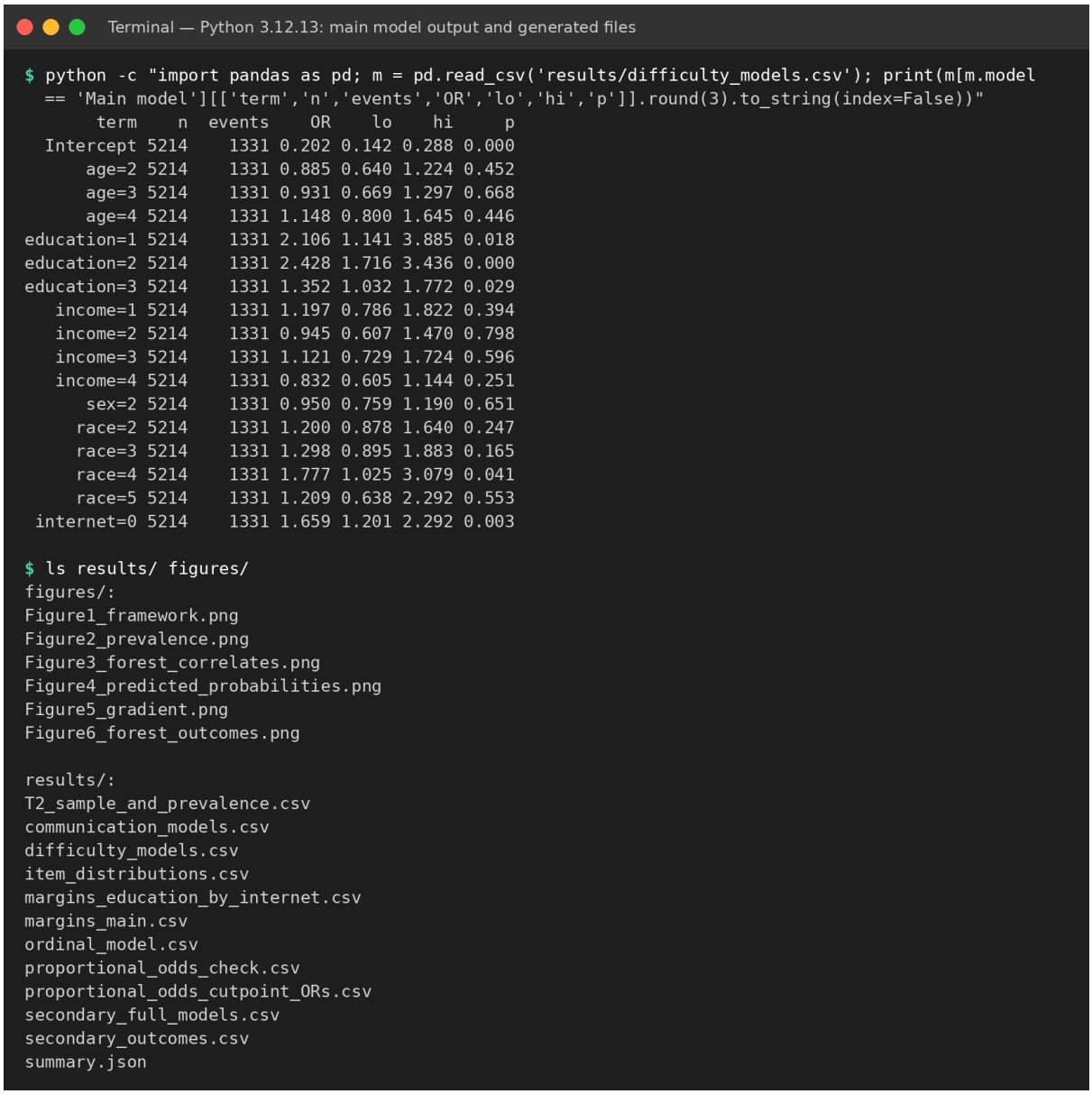


FIGURE B5. **MAIN ADJUSTED MODEL PRINTED FROM THE SAVED RESULTS FILE, AND THE RESULT TABLES AND FIGURES GENERATED BY THE RUN.**