Supplementary Material: Transforming Life Insurance Underwriting with LLM-driven Underwriting Workflows and Comorbidity Graphs

Anonymous Submission

I. Synthetic Dataset

The synthetic data generation process consisted of two separate yet linked processes: a) Generation of 28 Manual Chapters covering rateable medical impairments (e.g., diabetes), and b) Medical summaries of proposed applicants with rating (ground truth) using the chapters, as depicted in Figure A.1. Both types of synthetic data (chapters and summaries) were evaluated by human evaluators (underwriters) with domain expertise to validate quality using Synthetic Medical Data (SMD) scorecard as described in Section I-C and I-D [32]. The details are described below.

A. *Life Insurance Underwriting Manual Chapters Generation*

We use an LLM to generate Synthetic Underwriting Manual. In this process, we engage a medical domain expert who has significant knowledge in medical field and ask this expert to write prompt aiming at generating one specific medical chapter at a time. Each chapter is prompted to have the following elements: General Information, Underwriting Focus, Requirements, Rating Tables, and Additional Considerations. The prompt template used in this process is illustrated in Figure A.2.

Once these Underwriting Manual Chapters are generated, they are validated through human evaluations described in Section I-C.

B. *Life Insurance Medical Summaries and Underwriting Rating Generation*

We generate 91 Medical Summaries by engaging medical experts. We provided examples of anonymized description of proposed insured applicants. These are typically provided by the insurance agents to underwriters to obtain an estimate of underwriting ratings. These anonymized insurance applications were rephrased into standardized medical summaries using an LLM. The synthetic summaries were then reviewed further that whether there is any risk of re-identification does not remain. The two underwriter judges who evaluated these on compliance criteria of the SMD scorecard. These summaries are short text paragraphs describing the cases to be underwritten. They contain figurative information about the client (for example age and gender) and figurative lab results or description of impairments, and where appropriate dates and durations to establish a brief medical history. Accompanying each medical summary (case) is the ground truth rating. The rating generation process involves mapping information from the medical summary to the underwriting manual and applying guidelines to determine the appropriate rating. Figure 1 (from the main manuscript) illustrates this process using excerpts from the underwriting guidelines for Diabetes Mellitus and an example case for life insurance.

Step 1: Assess Each Impairment Individually. The underwriter (UW) begins by reviewing each impairment separately. For every impairment, a *base rating* is determined using factors such as the applicant's age and the severity of the condition. The underwriting manual provides tables that specify these base ratings. In the example shown in Figure 1 (from main manuscript), an applicant aged 57 with type 2 diabetes starts at a Standard rating (equivalent to +0). For comparison, a much younger applicant—say, age 17—with the same condition would receive a significantly higher base rating, such as +100, due to the higher risk associated with an early onset.

Once the base rating is established, the UW applies *credits* or *debits* based on further guidelines in the manual. These adjustments reflect how well the impairment is managed. For instance, the figure shows that good diabetes control can earn credits. An HbA1c value of 7.4 indicates good control, which might warrant a credit of -20, reducing the rating from +0 to -20. At this stage, if the manual specifies a decision such as *Postpone* (e.g., during the onset of an impairment or pending lab results) or *Decline* (e.g., severe impairment), that decision is considered final and replaces any numeric rating.

Step 2: Apply Co-morbidity Adjustments and Aggregate. After computing ratings for all relevant im-

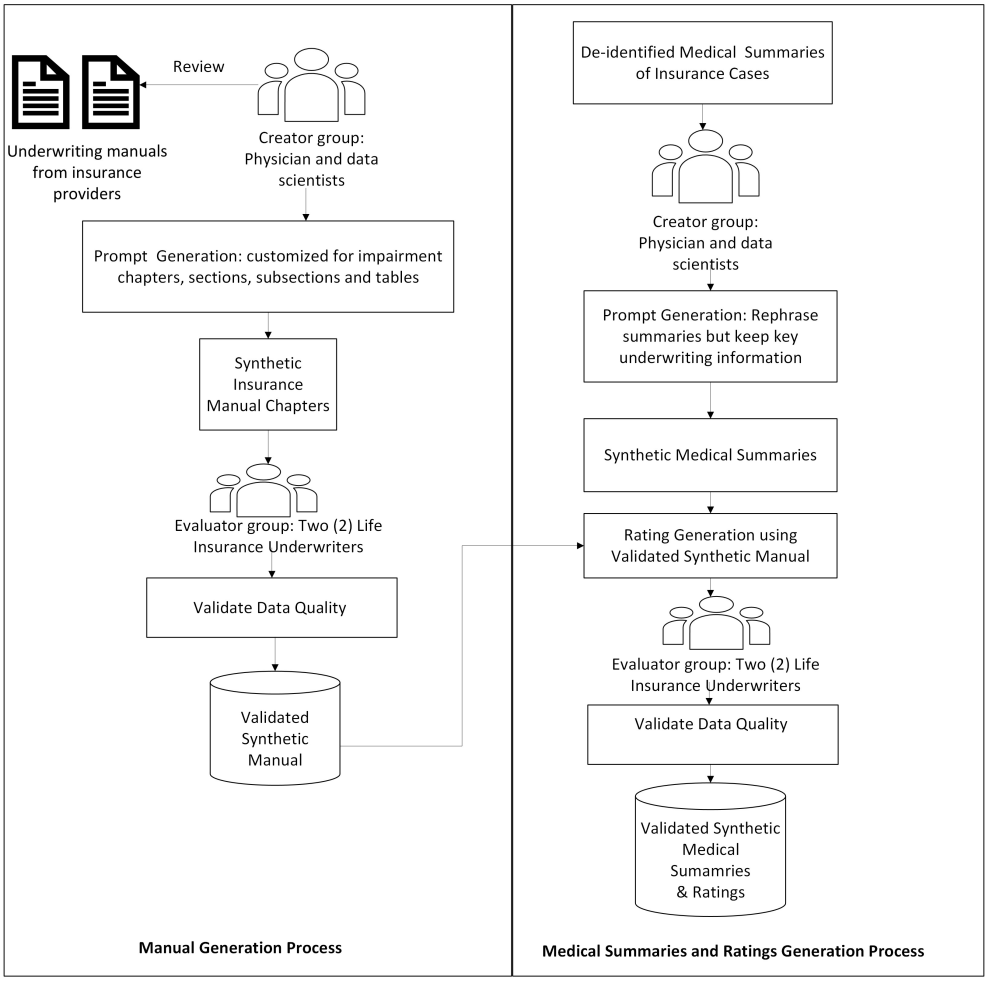
Review
Underwriting manuals from insurance providers
Creator group: Physician and data scientists
Prompt Generation: customized for impairment chapters, sections, subsections and tables
Synthetic Insurance Manual Chapters
Evaluator group: Two (2) Life Insurance Underwriters
Validate Data Quality
Validated Synthetic Manual
Manual Generation Process
De-identified Medical Summaries of Insurance Cases
Creator group: Physician and data scientists
Prompt Generation: Rephrase summaries but keep key underwriting information
Synthetic Medical Summaries
Rating Generation using Validated Synthetic Manual
Evaluator group: Two (2) Life Insurance Underwriters
Validate Data Quality
Validated Synthetic Medical Sumamries & Ratings
Medical Summaries and Ratings Generation Process

Fig. A.1. The process of synthetic data generation

pairments (e.g., diabetes, hypertension, high cholesterol), the UW applies co-morbidity guidelines. These rules specify how ratings interact when multiple impairments are present. For example, controlled hypertension may add +20 and high cholesterol +10 to the diabetes rating, changing it from -20 to +10. This adjusted value is referred to as the *co-morbidity-aware rating* for diabetes.

Finally, the underwriter (UW) aggregates ratings across impairments. When all impairments are co-morbid (i.e., related), the aggregation typically involves selecting the maximum rating among them. In cases where impairments form multiple co-morbidity sets, the UW first computes the maximum rating within each set and then sums these maxima. In the running example, diabetes, hypertension, and cholesterol form a co-morbid group, so the final rating is +10—the highest rating within that group.

Once these Medical Summaries and Accompanying Ground Truth Ratings are generated, they are validated through human evaluations described in Section I-D.

C. *Human Evaluation Criteria for Impairment Chapters*

Congruence (Alignment with Real Data)

- **1 – Low:** Synthetic chapter does not include common risk factors and related conditions (e.g., diabetes and BMI).
- **2 – Medium:** Most information related to common risk factors and related conditions is included.
- **3 – High:** All common and known risk factors and related conditions are included.

Coverage (Feature & Population Representativeness)

- **1 – Low:** Synthetic chapter has no information on relatively rare risk factors or related conditions (e.g., diabetes and gastroparesis).
- **2 – Medium:** Covers to some extent relatively rare risk factors or related conditions.

As the Director of Research of Underwriting in a life insurance company, write me chapters on the selected impairment ”[impairment name]” for the underwriting manual. The underwriting manual is used by junior underwriters to provide life underwriting rating. The underwriting manual consists of sections and chapters in which the impairments are grouped under each section. A typical chapter consists of following sub-chapters:

1) General Information: There are five sub-sub-chapters

1a) Definition and typical signs and symptoms if it is a medical condition.

1b) Risk factors affecting disease prognosis as well as protective factors that improve prognosis (e.g., compliance to taking medications or healthy behaviors)

1c) Classification of the severity of the impairment based on the validated US or Canadian criteria. Provide reference to the criteria and individual contributing factors and their cut-offs for each of the criteria.

1d) Diagnostic tests used for assessing disease severity.

1e) Treatments including generic names of medications and any other treatment available again referring to severity of impairments.

2) Underwriting Focus: This sub-chapter lists the factors from signs and symptoms, risk factors, diagnostic tests, and treatments that are known in medical literature to affect disease prognosis and disease severity. The purpose of this sub-chapter is to ensure that underwriters focus on these factors while considering underwriting. These factors also inform the underwriting tables.

3) Requirements: The third sub-chapter where applicable highlights the requirements for underwriting. For example, some of the risk factors are not readily available. Especially, if a specific diagnostic test report is required, it will be listed in this sub-chapter with cut-offs for severity with or without tabular format.

4) Rating: The fourth sub-chapter provides underwriting rates. The rating is typically provided as 0+ as standard rates, +50, +100, +150 or decline when it is 200+ depending on severity of the disease. Usually, the ratings must be provided by age, by sex or by smoking history based on life tables that are updated on internal life experience study. The published life tables indicate the mortality risk at different ages for each sex and by smoking status. The ratings are provided in tabular format with the possible variations for age, sex and smoking status.

5) Additional Considerations, where applicable: The sub-chapter has additional underwriting tables.

a) First it considers the co-morbid conditions and or factors that are expected to result in poor prognosis and hence results in higher rating than in the rating before. These factors are known to increase severity and are associated with hospitalizations and mortality. The risk factor-based tables have the severity of the key impairment on one axis and severity of the individual additional requirement or condition on the other axis.

b) It also considers a table of credits for protective factors. Usually the credits are -25, -50. The protective factors are compliance or having a protective treatment.

Format:

The output format of the ratings and additional considerations must be tabular and synthetic. For headings and subheadings in table use the html code provided below but do not extract cell values. I repeat do not extract cell values. I repeat that all values must be synthetic data.

Rating HTML Code:

””” One Shot Example of a typical table row and column headers are provided ”””

Additional Consideration Code:

””” One Shot Example of a typical table row and column headers are provided (if applicable) ”””

Fig. A.2. Underwriting Manual Chapter Generation Prompt

- **3 – High:** Includes full diversity including rare combinations of risk factors or related conditions.

Constraint (Adherence to Known Constraints)

- **1 – Low:** Violates clinical rules (e.g., male patients with pregnancy codes or values outside of possible ranges).
- **2 – Medium:** Mostly correct but occasional violations (e.g., BMI ¡10 or ¿80).
- **3 – High:** Fully respects clinical and technical constraints (e.g., all values biologically plausible).

Completeness (Absence of Missing or Incomplete Information)

- **1 – Low:** Many missing critical fields (e.g., diagnosis without treatment info).
- **2 – Medium:** Most fields present, but some secondary details missing.
- **3 – High:** All relevant fields complete for intended use case.

Compliance (Ethical, Clinical, and Regulatory Standards)

- **1 – Low:** No privacy safeguards; includes identifiable info.
- **2 – Medium:** De-identified but lacks adherence to any standards.
- **3 – High:** Fully anonymized, meets HIPAA or similar standards.

Comprehension (Interpretability & Transparency)

- **1 – Low:** Synthetic chapter is disorganized and not interpretable.
- **2 – Medium:** Overall understandable but can be improved.
- **3 – High:** Fully understandable and organized.

D. *Human Evaluation Criteria for Medical Summaries*

Congruence (Alignment with Real Data)

- **1 – Low:** Synthetic medical summary does not include key information from the original summary.
- **2 – Medium:** Synthetic medical summary includes most of the key information from the original summary.
- **3 – High:** Synthetic medical summary includes all the key information from the original summary.

Constraint (Adherence to Known Constraints)

- **1 – Low:** Violates clinical rules (e.g., male patients with pregnancy codes or values outside of possible ranges).
- **2 – Medium:** Mostly correct but occasional violations (e.g., BMI < 10 or > 80).
- **3 – High:** Fully respects clinical and technical constraints (e.g., all values biologically plausible).

Compliance (Ethical, Clinical, and Regulatory Standards)

- **1 – Low:** No privacy safeguards; includes identifiable info.
- **2 – Medium:** De-identified but lacks adherence to any standards.
- **3 – High:** Fully anonymized, meets HIPAA or similar standards.

Comprehension (Interpretability & Transparency)

- **1 – Low:** Synthetic medical summary is disorganized and not interpretable.
- **2 – Medium:** Synthetic medical summary is overall understandable but can be improved.
- **3 – High:** Synthetic medical summary is fully understandable and organized.

E. Statistical Analyses of Human Evaluation Metrics

To assess inter-rater reliability between two independent underwriters, we computed Cohen's kappa (κ) and the Intraclass Correlation Coefficient (ICC3). Cohen's κ measures agreement beyond chance and is defined as:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where P_o is the observed agreement and P_e is the expected agreement by chance. Values of κ range from -1 to 1, with values above 0.6 generally considered substantial and values above 0.8 indicating near-perfect agreement [33].

The ICC3 statistic, based on a two-way mixed-effects model, evaluates consistency among raters and is calculated as:

$$ICC = \frac{MS_R - MS_E}{MS_R + (k-1)MS_E}$$

where MS_R is the mean square for subjects, MS_E is the mean square for error, and k is the number of raters. ICC values above 0.75 indicate good reliability, and values above 0.9 are considered excellent.

In our evaluation, both metrics demonstrated strong agreement across most criteria. For medical chapters, Cohen's kappa values ranged from 0.65 to 0.84 (see Table A.1), and ICC3 values were consistently above 0.75 (see Table A.2), indicating substantial to good reliability. For medical summaries, Cohen's kappa values ranged from 0.66 to 0.89, with ICC3 reaching as high as 0.91 for comprehension, reflecting excellent agreement. These results confirm that the SMD Scorecard ratings are reliably reproducible across independent evaluators.

TABLE A.1
EVALUATION METRICS FOR MEDICAL CHAPTERS

Criterion	**Average Score**	**ICC3**	**Cohen's Kappa**
Congruence	2.85	0.7886	0.8432
Constraint	2.88	0.8811	0.8170
Compliance	2.91	0.7874	0.7812
Comprehension	2.95	0.6582	0.6500
Coverage	2.91	0.8525	0.7358
Completeness	2.95	0.6582	0.6500

TABLE A.2
EVALUATION METRICS FOR MEDICAL SUMMARIES

Criterion	**Average Score**	**ICC3**	**Cohen's Kappa**
Congruence	2.89	0.6669	0.8903
Constraint	2.96	0.6089	0.7069
Compliance	2.99	0.6645	0.6623
Comprehension	2.97	0.9079	0.7964
Coverage	N/A	N/A	N/A
Completeness	N/A	N/A	N/A

II. Statement of Reproducibility

The code and data used in this paper have been uploaded to the paper review system. The steps to reproduce the paper's results are:

- Download the code repo (Data is stored inside code repo)
- Open the repo in VScode Editor
- Follow the guide on ReadMe.md