The sample calculation of the CPLIP model in the benchmark stage

This is a **retrospective cohort study** of patients with oral potentially malignant disorders (OPMDs) with a **36-month** follow-up window from the index diagnostic biopsy/WSI. We use a CPLIP-based risk model to classify patients **at baseline** as high risk or low risk.

The outcome is malignant transformation to oral cancer within 36 months, confirmed by biopsy. For validation, the 36-month biopsy-confirmed outcome serves as the **ground truth**, and we evaluate performance at the **patient level**.

	No Progression	Progression
Low-Risk		
High-Risk		

Overall, the malignant transformation rate across all OPMD groups was 7.9% (99% confidence interval [CI], 4.9%-11.5%)(Iocca et al., 2020). Splits are patient-level (no patient appears in more than one split).

How many patients to collect for the benchmark
Use the planning rule: patients = target events ÷ event rate

· To target ≈100 events (a solid benchmark for pooled 5-fold CV):
N=100/0.079=1,265.8→ ~**1,300 patients** (expected ≈103 events).
This gives ≈21 events per fold in 5-fold CV.

· Add buffer for missing outcomes (e.g., **10% loss** to follow-up):
1,266/0.90≈ ~1,400-1,450 patients.

· If I want to be conservative and plan against the lower bound of the rate (4.9%):
100/0.049≈2,041 → ~2,250 if I also allow 10% loss.

· Minimal pilot (pipeline shake-down): ~650 patients (~50 events).

· Tighter AUC CIs (≈±0.05): ~1,600–1,800 patients (≈125-140 events).

・ Above the cutoff → label High-risk

・ Below the cutoff → label Low-risk

Assume **N=1,300 patients** and 7.9% will transform.
Choose a cutoff that gives Sensitivity ≥ 0.90
→ expect about **570 high-risk** / **730 low-risk** (≈44% / 56%).

How to do this in 5-fold CV
In the 5-fold CV steps, the threshold is chosen inside the training data via inner validation (nested CV) to avoid leakage; then apply it to the held-out fold.
①Split patients into 5 folds (patient-level, stratified).
②For each fold, train on the other 4 folds.
③Choose the cutoff inside the training data (e.g., Sensitivity ≥ 0.90, then maximize Specificity).
④Apply that cutoff to the held-out fold; record who is high/low.
⑤Pool all 5 held-out predictions to report final counts and metrics.

5-fold CV is better than 70:15:15
・More test signal. With N=1,300 and 7.9% events:
70:15:15 ⟶ test has ~195 people ⟶ ~15 events.
5-fold CV ⟶ everyone is tested once ⟶ ~1,300 people ⟶ ~103 events.
More events = tighter, more trustworthy metrics.
· Better training. Each CV model trains on ~80% of the data (vs 70%), so it learns more.
· Less luck. CV averages over 5 different test folds, not one lucky/unlucky split.
· Practicality. To get ~100 test events with a 15% test set, I'd need ~8,440 total patients. CV gives ~100 test events now by pooling the 5 folds.

Iocca, O., Sollecito, T. P., Alawi, F., Weinstein, G. S., Newman, J. G., De Virgilio, A., Di Maio, P., Spriano, G., Pardiñas López, S., & Shanti, R. M. (2020). Potentially malignant disorders of the oral cavity and oral dysplasia: A systematic review and meta-analysis of malignant transformation rate by subtype. *Head & neck*, *42*(3), 539-555.