

Oral health analytics: a ClinicalSpark demonstration

Title: Masticatory dysfunction, tooth loss and prosthetic rehabilitation in relation to cardiometabolic outcomes: a ClinicalSpark demonstration

Running Title: Oral health analytics: a ClinicalSpark demonstration

Authors: Antigravity Science & Clinical Analytics Group

Oral health analytics: a ClinicalSpark demonstration

ABSTRACT

Background: Oral function, tooth retention and prosthetic rehabilitation are distinct dimensions of oral health. This ClinicalSpark demonstration illustrates their analysis in relation to cardiometabolic outcomes; it is not independently validated clinical research.

Methods: The demonstration describes linkage of examination, registry, medical-claims and dental records for a nominal population of 3,420,777 adults. Supplied outputs comprise descriptive comparisons, Cox regression, propensity score matching on age, sex, body mass index and systolic blood pressure, a metabolic decomposition, and gradient boosted decision tree prediction of a five-year cardiovascular endpoint. No new participant-level analysis was performed.

Results: The matched analysis comprised 585,518 pairs and reported a mortality hazard ratio of 1.412 (95% confidence interval, 1.385–1.440) for impaired chewing. A separate prosthetic rehabilitation coefficient was 0.428 (0.368–0.498). Reported five-year survival was 97.4%, 94.1% and 84.2% with ≥ 20 , 10–19 and < 10 remaining teeth, respectively. The anatomical output assigned mortality hazard ratios of 4.846 to molar loss and 2.074 to anterior-only loss; their quotient is not a directly estimated contrast. The decomposition assigned 31.4% to glycated hemoglobin and systolic blood pressure components. The prediction output reported an area under the receiver operating characteristic curve of 0.842 (0.836–0.848). These illustrative quantities do not establish clinical validity, causal mediation or treatment benefit.

Conclusions: The demonstration brings oral function, tooth-level information and rehabilitation into a common analytical framework. Its descriptive, etiological and predictive questions remain distinct, and its estimates do not provide evidence for clinical decisions.

Keywords: Mastication; Tooth loss; Dental prosthesis; Cardiovascular diseases; Mortality; Propensity score; Machine learning

1. INTRODUCTION

Masticatory difficulty can be studied as a patient-reported functional problem, alongside anatomical measures such as remaining teeth and posterior occlusal support. These measures capture different aspects of oral health and need not identify the same people. In a longitudinal Singapore cohort, poorer perceived masticatory function was associated with all-cause mortality; the investigators also evaluated hypothetical interventions using longitudinal causal methods. That study supports the relevance of oral function to healthy aging, while its hypothetical interventions remain distinct from observed effects of prosthodontic treatment.¹

Cardiometabolic associations provide a complementary rationale. The Suita Study linked lower measured masticatory performance to subsequent metabolic syndrome in men, whereas comparable associations were not statistically significant in women. A cross-sectional analysis of the Aichi Workers' Cohort Study found an association between higher fasting glucose and self-reported chewing difficulty. Together, these studies motivate investigation across functional and metabolic domains, but they differ in exposure measurement, population and temporal design. They do not establish a single direction of causation between chewing difficulty and dysglycemia.^{2,3}

Linked examination and claims records offer a setting in which oral function, treatment history and systemic outcomes can be represented together. The scientific questions nevertheless remain distinct: whether oral dysfunction is associated with an outcome, whether rehabilitation changes that outcome, whether specified variables mediate an effect, and whether oral measures improve prediction. Neither a large record count nor the use of several analytical methods makes these questions interchangeable.

24 This manuscript presents an existing ClinicalSpark Studio demonstration organized
25 around those four questions. It describes the supplied linkage framework and illustrative
26 outputs for chewing function, prosthetic rehabilitation, tooth location and remaining
27 dentition, followed by matching, mediation and prediction analyses. Its contribution is an
28 integrated analytical example, with quantitative results retained at their demonstration
29 evidence level.

30 2. MATERIALS AND METHODS

31 2.1. Demonstration design and analytical material

32 The material comprised a supplied ClinicalSpark Studio manuscript containing a
33 workflow description and aggregate numerical outputs. ClinicalSpark Studio was
34 described as using an Apache Spark core. No participant-level records were analyzed for
35 the present manuscript, and the reported estimates were not independently regenerated.
36 The numerical results therefore represent the existing demonstration, rather than a newly
37 established population cohort or a validation study. The provenance of the underlying
38 demonstration records does not permit classification here as either independently
39 observed clinical data or a defined simulation.

40 The described linkage combines 5,305,400 health examination records
41 (exam_interview_processed), 13,114,600 registry records (tekiyo),
42 2,257,364,242 diagnosis records (receipt_diseases), 498,777,327 dental claim
43 items (receipt_dental_practice) and 1,421,861,940 tooth-level disease records
44 (receipt_tooth_type_diseases). The stated analysis population is 3,420,777
45 individuals. These are different record units: source-table counts are not counts of
46 independent participants, and their sum does not define the analysis population.

2.2. Oral function, rehabilitation and dentition

The workflow classifies chewing function using `sosyaku_code`: a value of 1 denotes the control category and values ≥ 2 denote impaired chewing. Prosthetic rehabilitation is identified by claim items described as removable dentures, fixed bridges or implants. The supplied groups comprise 2,816,907 controls, 585,518 people with impaired chewing and rehabilitation, and 18,352 with impaired chewing and no recorded rehabilitation. Absence of a recorded claim is the operational distinction; it is not proof that rehabilitation was never received.

Remaining dentition is represented by three categories: ≥ 20 , 10–19 and < 10 teeth. The last category is termed severe tooth loss, because it includes people with remaining teeth and is not synonymous with edentulism. Tooth-site information (`shishiki`) identifies loss or disease of posterior molars, coded 16–18, 26–28, 36–38 and 46–48. The anterior-only category restricts loss or disease to incisors and canines, coded 11–13, 21–23, 31–33 and 41–43. The anatomical description combines loss or disease at these sites; the resulting flags are therefore not established measurements of missing teeth alone.

The source assigns the same three group sizes to rehabilitation and dentition categories and reuses two sizes in the anatomical analysis. These summaries are retained as separate demonstration classifications, without assuming participant-level equivalence, independence or a verified cross-tabulation. The matched analysis is likewise reported for its stated 585,518 exposed participants rather than equated with all 603,870 participants classified as having impaired chewing.

2.3. Outcome definitions

The demonstration defines diabetes by glycated hemoglobin (HbA1c) $\geq 6.5\%$ or fasting glucose ≥ 126 mg/dL, and hypertension by systolic blood pressure (SBP) ≥ 140 mmHg or diastolic blood pressure ≥ 90 mmHg. These are operational examination thresholds. Diabetes proportions reported by tooth site are prevalence summaries and are not incident risks.

The mortality flag is `tenki_kbn_code = 3` in the diagnosis table. This is treated as the demonstration's claims-based mortality proxy, without asserting complete ascertainment of all-cause mortality. The five-year cardiovascular endpoint (`cvd_event`) is described using ICD-10 groups I20–I25, I60–I69 and I50. These groups cover ischemic heart disease, cerebrovascular disease and heart failure; they are broader than acute myocardial infarction and acute stroke alone. The supplied incident-event label does not, by itself, establish exclusion of prevalent disease or a validated event algorithm.

2.4. Descriptive, survival and matched analyses

Continuous characteristics are reported as means and standard deviations. The supplied survival analyses use Cox models and Kaplan–Meier estimates, with nominal confidence intervals and P values retained as demonstration outputs. The prosthetic Cox summary contains terms for unrestored impairment, rehabilitation, SBP and body mass index (BMI). Because complete model coding and reference levels are not established, those coefficients are not interpreted as a verified comparison of rehabilitated and unrehabilitated patients. Crude mortality percentages and time-to-event estimates are treated as different quantities.

For propensity score matching, the described exposure model is logistic regression with age, sex, BMI and SBP. Matching is specified as 1:1 nearest-neighbor matching without replacement, using a caliper of 0.2 standard deviations of the logit propensity score. The supplied balance summaries use standardized mean differences (SMDs), with values below 0.05 reported after matching. Matching addresses the included covariates; it is not evidence that all confounding has been removed. The matched mortality risk ratio and hazard ratio are retained separately because they describe different outcome scales.

2.5. Mediation and prediction components

The mediation component (`causal_dag`) describes an additive decomposition of a total effect into a direct component and indirect components through HbA1c and SBP. The reported mediated proportion divides the summed indirect components by the total effect. In this demonstration, the effect scale and temporal ordering do not establish a causal estimand. The decomposition is consequently presented as a numerical illustration; the direct component is not assigned to an observed inflammatory or neural mechanism.

The prediction component is a gradient boosted decision tree model for the stated five-year cardiovascular endpoint. Its supplied metrics comprise the area under the receiver operating characteristic curve (ROC-AUC), sensitivity, specificity and F1 score, together with relative feature importance. These are reported model outputs. They do not establish performance in an independent test population or quantify the clinical benefit of acting on a prediction. Feature importance describes the fitted model's allocation of importance, rather than a causal contribution to cardiovascular disease.

3. RESULTS

3.1. Descriptive patterns and rehabilitation-related outputs

The three rehabilitation groups sum to the stated demonstration population of 3,420,777. Mean HbA1c and SBP were higher in both impaired-chewing groups than in controls, while mean BMI was similar in controls and the rehabilitation group and higher in the group without recorded rehabilitation (Table 1). These descriptive summaries do not isolate the effect of oral function or treatment.

Table 1. Supplied demonstration characteristics by chewing and rehabilitation category.

Characteristic	Control	Impaired, rehabilitated	Impaired, no rehabilitation
Group size	2,816,907	585,518	18,352
HbA1c, %	5.63 ± 0.62	5.73 ± 0.69	5.80 ± 1.15
SBP, mmHg	124.79 ± 17.60	127.43 ± 17.80	127.61 ± 18.06
Diastolic BP, mmHg	74.97 ± 11.38	75.22 ± 11.33	75.80 ± 11.88
BMI, kg/m ²	23.03 ± 3.62	23.00 ± 3.64	23.90 ± 4.24

Values other than group size are mean ± standard deviation. Rehabilitation denotes recorded prosthetic claims in the supplied classification. BP, blood pressure; BMI, body mass index; HbA1c, glycated hemoglobin; SBP, systolic blood pressure.

The supplied crude mortality proportions were 3.673% for unrestored impairment and 0.683% for controls. The accompanying log-rank statistic was $\chi^2 = 5,638.72$, with $P < 0.0001$. In the separate Cox summary, the unrestored-impairment coefficient was 4.846 (95% confidence interval [CI], 4.176–5.623), while the rehabilitation coefficient was 0.428 (0.368–0.498; Table 2). The latter is mathematically 57.2% below a hazard ratio of 1; it does not demonstrate that rehabilitation reduced mortality by 57.2%.

Table 2. Supplied demonstration Cox model coefficients for the rehabilitation analysis.

Term	Hazard ratio (95% CI)	z statistic	P value
Unrestored impaired chewing	4.846 (4.176–5.623)	20.84	6.38×10^{-96}
Prosthetic rehabilitation	0.428 (0.368–0.498)	–11.00	5.25×10^{-28}
SBP, per mmHg	1.013 (1.012–1.013)	45.92	<0.0001
BMI, per kg/m ²	0.947 (0.941–0.953)	–16.82	1.02×10^{-63}

Coefficients and inferential quantities are reproduced as demonstration outputs. Complete model coding and categorical reference levels are not established; the rehabilitation term is not a verified head-to-head treatment effect. CI, confidence interval.

3.2. Matched mortality comparison

The matched summary comprised 585,518 exposed participants and 585,518 controls, for a total of 1,171,036. Reported post-match SMDs ranged from 0.000 to 0.002 for the four matching covariates (Table 3). The supplied mortality risk ratio was 1.284 (95% CI, 1.268 –1.300), and the hazard ratio was 1.412 (1.385–1.440), both with $P < 0.0001$. These outputs show the form of a matched analysis; the displayed SMDs alone do not validate cohort selection or the mortality model.

Table 3. Supplied propensity score matching summaries.

Covariate	Before: control	Before: impaired	SMD before	After: both groups	SMD after
Age, years	54.2 ± 10.1	58.4 ± 9.2	0.384	58.4 ± 9.2	0.002
Male, %	51.8	54.2	0.112	54.2	0.000
SBP, mmHg	124.8 ± 17.6	127.4 ± 17.8	0.151	127.4 ± 17.8	0.001
BMI, kg/m ²	23.03 ± 3.6	23.01 ± 3.6	0.003	23.01 ± 3.6	0.000

Continuous values are mean ± standard deviation. The matched groups each contain 585,518 participants as reported. "After: both groups" contains the identical rounded summaries supplied for the two groups. SMDs are source-reported values, not recalculations from those rounded summaries. SMD, standardized mean difference.

3.3. Anatomical and remaining-dentition outputs

The tooth-site summary reported diabetes prevalence of 13.95% in the molar category and 10.16% in the anterior category (Table 4). Its supplied ratios have unspecified

reference populations and therefore do not quantify a validated molar-versus-anterior comparison. The quotient of the two mortality hazard ratios, 4.846/2.074, is approximately 2.34. Without a directly estimated contrast and its uncertainty, that quotient cannot establish the magnitude or statistical significance of a difference between tooth sites.

Table 4. Supplied anatomical outputs for diabetes prevalence and mortality.

Output	Molar loss/disease	Anterior-only loss/disease
Diabetes count/group size	2,560/18,352	59,485/585,518
Diabetes prevalence, %	13.95	10.16
Reported ratio (95% CI)	1.752 (1.690–1.817)	1.355 (1.344–1.367)
Odds ratio (95% CI)	1.874 (1.797–1.955)	1.395 (1.382–1.409)
Diabetes P value	<0.0001	<0.0001
Mortality hazard ratio (95% CI)	4.846 (4.176–5.623)	2.074 (1.979–2.173)
Mortality P value	6.38×10^{-96}	1.73×10^{-205}

The source labels the diabetes ratio "risk ratio," although the displayed outcome is prevalent diabetes. Ratio reference populations and adjustment specifications are not established. The P values are supplied for the separate outputs and are not tests of a molar-versus-anterior contrast. Anatomical categories include loss or disease, as described in the workflow.

Five-year survival declined across the three supplied remaining-teeth categories, from 97.4% with ≥ 20 teeth to 94.1% with 10–19 teeth and 84.2% with < 10 teeth (Table 5). The difference between the highest and lowest categories is 13.2 percentage points. These Kaplan–Meier summaries describe a different quantity from the crude mortality percentages and do not establish that tooth loss itself caused the survival gradient.

Table 5. Supplied five-year survival outputs by remaining dentition.

Remaining teeth	Group size	Survival, % (95% CI)	Hazard ratio (95% CI)
≥20	2,816,907	97.4 (97.3–97.5)	1.000 (reference)
10–19	585,518	94.1 (93.9–94.3)	1.684 (1.642–1.727)
<10	18,352	84.2 (83.4–85.0)	4.846 (4.176–5.623)

Survival values are labeled Kaplan–Meier estimates in the demonstration. $P < 0.0001$ was supplied for each nonreference row as a log-rank result. The row-specific comparison structure, follow-up distribution and Cox adjustment set are not established. The <10-teeth category denotes severe tooth loss, not complete edentulism.

3.4. Illustrative metabolic decomposition

The supplied total coefficient was 0.0412, comprising a direct component of 0.0283 and indirect components of 0.0071 through HbA1c and 0.0058 through SBP (Table 6). The reported component percentages were 68.6%, 17.2% and 14.2%, respectively, with a combined mediated proportion of 31.4%. These percentages are retained at the precision supplied. They describe the demonstration decomposition, rather than the proportion of clinical cardiovascular events caused through either mediator. No measured inflammatory or trigeminal pathway is identified by the direct component.

Table 6. Supplied numerical decomposition for the cardiovascular endpoint.

Component	Coefficient	Reported contribution, %
Total	0.0412	100.0
Direct	0.0283	68.6
Indirect through HbA1c	0.0071	17.2
Indirect through SBP	0.0058	14.2
Combined indirect	0.0129	31.4

All rows were supplied with $P < 0.0001$. Coefficients and percentages are reproduced at source precision; the coefficient scale and causal identification conditions are not established. The combined indirect row sums the two preceding components and is not an additional pathway.

3.5. Cardiovascular prediction output

The GBDT summary reported ROC-AUC 0.842 (95% CI, 0.836–0.848), sensitivity 78.4%, specificity 76.1% and F1 score 0.749. Molar loss and chewing difficulty together

accounted for 48.1% of the supplied feature-importance allocation (Table 7). This allocation is internal to the reported model and does not show that oral features explain 48.1% of cardiovascular risk or add that amount of predictive value beyond conventional risk factors.

Table 7. Supplied relative feature importance in the cardiovascular prediction model.

Feature	Importance, %
Molar loss flag	28.4
Systolic blood pressure	24.1
Chewing difficulty	19.7
HbA1c	15.2
Body mass index	7.8
Prosthetic rehabilitation flag	4.8

The importance values sum to 100.0%. The model's importance definition, evaluation population and operating threshold are not established. Values are descriptive model outputs, not causal effect sizes. GBDT, gradient boosted decision tree; ROC-AUC, area under the receiver operating characteristic curve.

4. DISCUSSION

This ClinicalSpark demonstration places functional oral symptoms, dental anatomy and prosthetic claims within a single account of cardiometabolic and survival analyses. Its supplied outputs show gradients in metabolic measurements and survival, a matched mortality association, a numerical mediation decomposition and a cardiovascular prediction summary. The supported contribution is the organization of these analytical questions. The demonstration does not establish that tooth loss is an independent causal driver of cardiovascular disease or that prosthetic rehabilitation prolongs life.

The distinction between functional symptoms and anatomical exposure is central to interpretation. A patient may report chewing difficulty despite retaining teeth, and a prosthetic claim does not measure the functional performance of the restoration.

Oral health analytics: a ClinicalSpark demonstration

205 Likewise, a tooth-level disease code does not necessarily indicate loss of that tooth.
206 Keeping these constructs separate is necessary for interpreting both the apparent survival
207 gradient and the model's allocation of importance to molar loss and chewing difficulty.
208 Repeated group sizes in the illustrative summaries cannot substitute for joint
209 measurement of these constructs.

210 The external literature supports continued investigation of oral function while also
211 showing why study design matters. Tay and colleagues assessed changing perceived
212 masticatory function and mortality using longitudinal data, whereas the present material
213 provides only aggregate demonstration summaries. Fushida and colleagues used an
214 objective chewing-performance measure and found sex-specific associations with
215 metabolic syndrome. Hamrah and colleagues studied fasting glucose and chewing
216 difficulty cross-sectionally. These differences prevent direct validation of the
217 demonstration by numerical or directional agreement with those studies.¹⁻³

218 The rehabilitation coefficient illustrates a particularly consequential distinction
219 between association and intervention. A hazard ratio below 1 can arise in a treatment-
220 associated comparison without measuring the effect of assigning treatment. Treatment
221 eligibility, health status and access to care can influence which patients receive
222 prostheses; treatment timing must also be aligned with the start of follow-up. The
223 illustrative coefficient therefore cannot support reimbursement or cardiovascular
224 prevention recommendations based on a claimed mortality benefit. The same reasoning
225 applies to the matched estimate: balance on four specified covariates does not, by itself,
226 turn a comparison into an intervention effect.

227 The metabolic decomposition and prediction model address different scientific aims.
228 An additive decomposition has a causal interpretation only with a defined effect scale,

temporal structure and identifying assumptions. Its residual direct component cannot be equated with inflammation, bacteremia or neural signaling without measurements that distinguish those explanations. Prediction instead concerns performance for a defined target population and time horizon. An AUC and a feature-importance ranking alone do not establish calibration, transportability or the value of changing clinical decisions.

Taken together, the demonstration provides a structured example for investigating oral function and systemic outcomes through linked records. Descriptive associations, treatment contrasts, mediation and prediction should retain their distinct interpretations. The supplied numerical outputs are useful for illustrating those distinctions, but they do not provide a basis for clinical efficacy claims or changes to care.

REFERENCES

1. Tay JRH, Cooray U, Chan A, Tonetti MS, Nascimento GG, Peres MA. Perceived Masticatory Function and Mortality: A Causal Study. *J Dent Res*. 2026;105(1):95-102. <http://dx.doi.org/10.1177/00220345251363851>
2. Fushida S, Kosaka T, Nakai M, Kida M, Nokubi T, Kokubo Y, et al. Lower Masticatory Performance Is a Risk for the Development of the Metabolic Syndrome: The Suita Study. *Front. Cardiovasc. Med*. 2021;8:752667. <http://dx.doi.org/10.3389/fcvm.2021.752667>
3. Hamrah MH, Song Z, Hong Y, Akter T, Nuamah HG, Gondo N, et al. Association between fasting blood glucose level and difficulty with chewing: the Aichi Workers' Cohort Study. *Environ Health Prev Med*. 2025;30(0):95. <http://dx.doi.org/10.1265/ehpm.25-00284>