

Supplementary Material of “*Quality of life in head and neck cancer survivors: the Big Data for Quality of Life study*”

1 **Mauricio Moreira-Soares^{1*}, Erlend I. F. Fossen¹, Katherine J. Taylor², Susanne Singer²,**
2 **Katrina Hurley^{3a}, Steve Thomas^{3b}, Miranda Pring^{3b}, Andrew Ness^{3b}, Stefano Cavalieri^{4,5},**
3 **Claudia Vener⁶, Laura Lopez-Perez⁷, Maria Fernanda Cabrera-Umpierrez⁷, Giuseppe Fico⁷,**
4 **Arnoldo Frigessi^{1,8}, Lisa Licitra^{4,5}, Marissa LeBlanc^{1,9}, on behalf of the BD4QoL consortium**

5 ¹ Oslo Centre for Biostatistics and Epidemiology, University of Oslo, Oslo, Norway

6 ² Institute of Medical Biostatistics, Epidemiology, and Informatics, University Medical Centre
7 Mainz, Germany

8 ³ Bristol Head & Neck 5000 Study, 3a University Hospitals Bristol and Weston NHS Foundation
9 Trust, Bristol, UK, 3b Bristol Dental School, University of Bristol, United Kingdom

10 ⁴ Head and Neck Medical Oncology Department, Fondazione IRCCS Istituto Nazionale dei Tumori,
11 Milan, Italy

12 ⁵ Department of Oncology and Hemato-oncology, University of Milan, Italy

13 ⁶ Epidemiology and Prevention Unit, Fondazione IRCCS Istituto Nazionale dei Tumori di Milano,
14 Milan, Italy

15 ⁷ Universidad Politécnica de Madrid-Life Supporting Technologies Research Group, ETSIT, Madrid,
16 Spain

17 ⁸ Oslo Centre for Biostatistics and Epidemiology, Oslo University Hospital, Oslo, Norway

18 ⁹ Norwegian Institute of Public Health, Oslo, Norway

19 *** Correspondence:**

20 Corresponding Author

21 m.m.soares@medisin.uio.no

22

23

24

25

26

27

28

29

30

31 1 Supplementary Material

32 1.1 Data harmonization

33 Differences and commonalities in the studies variables

34 We observed high heterogeneity with respect to variables collected across studies. For instance, the
35 UMM data contain a broad range of variables with detailed information about treatment tolerance,
36 clinical devices used, welfare status such as sick leave and EU pension applications, among many
37 other demographic and socioeconomic indicators. On the other hand, the INT data focus mainly on
38 clinical information. The UoB data present a mix of clinical information and QoL data with several
39 different questionnaires. We integrated basic clinical characteristics, demographics, and two quality-
40 of-life questionnaires that were present in all datasets, the EORTC QLQ-C30 and the head and neck
41 module EORTC QLQ-H&N35. For variables that were not completely equivalent and presented
42 different levels of details, e.g., education level, household income, and marital status, we reduced
43 their granularity to allow comparable measurements.

44 Time-to-event data

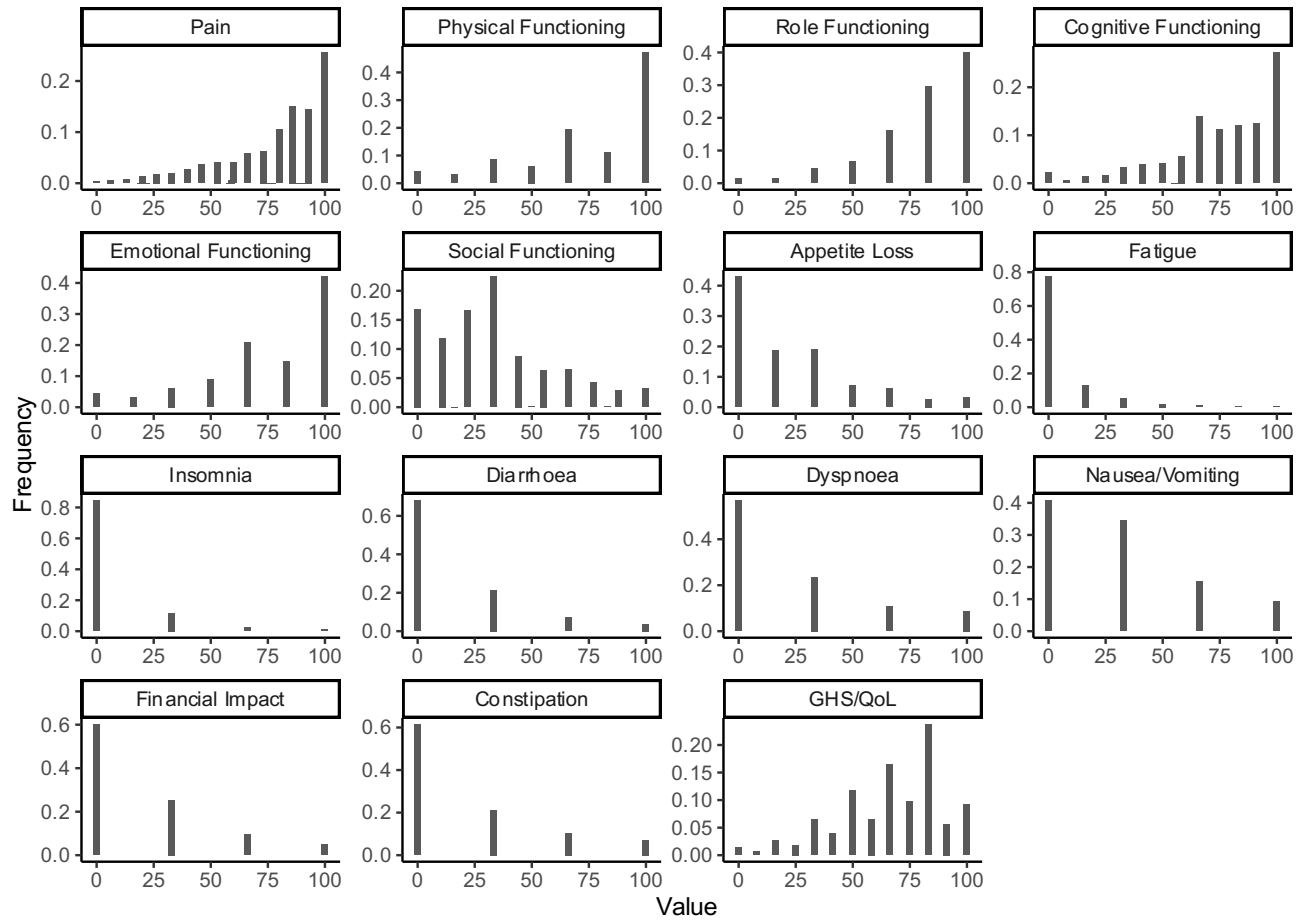
45 The UMM2 and UMM3 studies were not intended to model survival, so only interval-censored data
46 with right censoring are available. Patients' statuses were checked at scheduled timepoints, but times
47 of death were not recorded. All patient statuses were measured at baseline, so left censoring is not
48 present in the data. UMM1 provided vital information for 83% of the subjects (N=181) from study
49 enrollment (before treatment) up to 12 years. We found no significant differences between the groups
50 with measured and missing vital information age, sex, tumor location, tumor stage, income, and
51 education level, using a univariate logistic regression model approach where the response variable is
52 the binary indicator "*missing vital information*" (yes/no). Therefore, we assume that the vital
53 information is missing completely at random (MCAR) due to the data collection procedure, which
54 involved retrieving vital information for each patient directly from clinicians many years after the
55 study's conclusion. INT provided the time of death/censoring from study enrollment (before
56 treatment) up to two years. UoB provided survival data with time of death/censoring from the time of
57 study enrollment (before treatment). Participants' statuses were checked every 6 months from UK
58 registry data up to 6 years after study enrollment. The survival data for INT, UMM1, and UoB were
59 aligned and shifted to the BD4QoL baseline (after treatment), which consisted of a 6-month shift for
60 INT, 4 months for UMM1, and 12 months for UoB.

61

62 2 Supplementary Figures and Tables

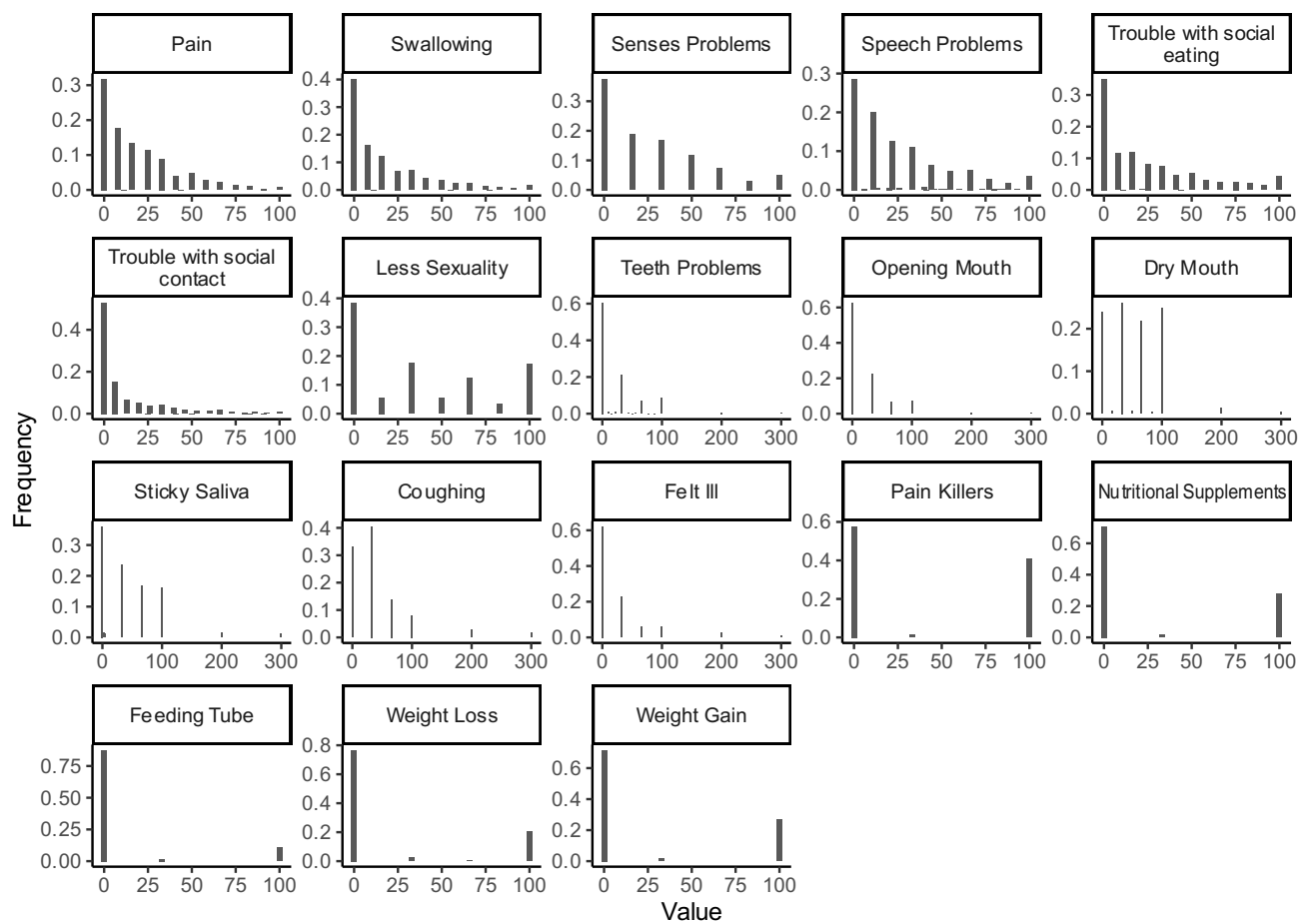
63

64 2.1 Supplementary Figures



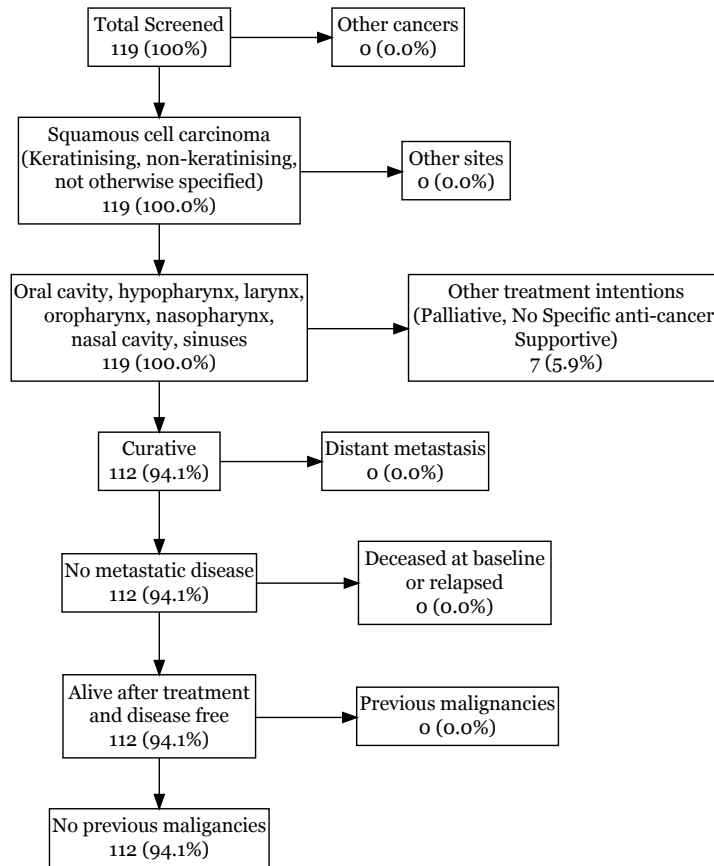
74

75 **Figure S2 EORTC QLQ-C30 functional and symptoms scales histograms.**



76

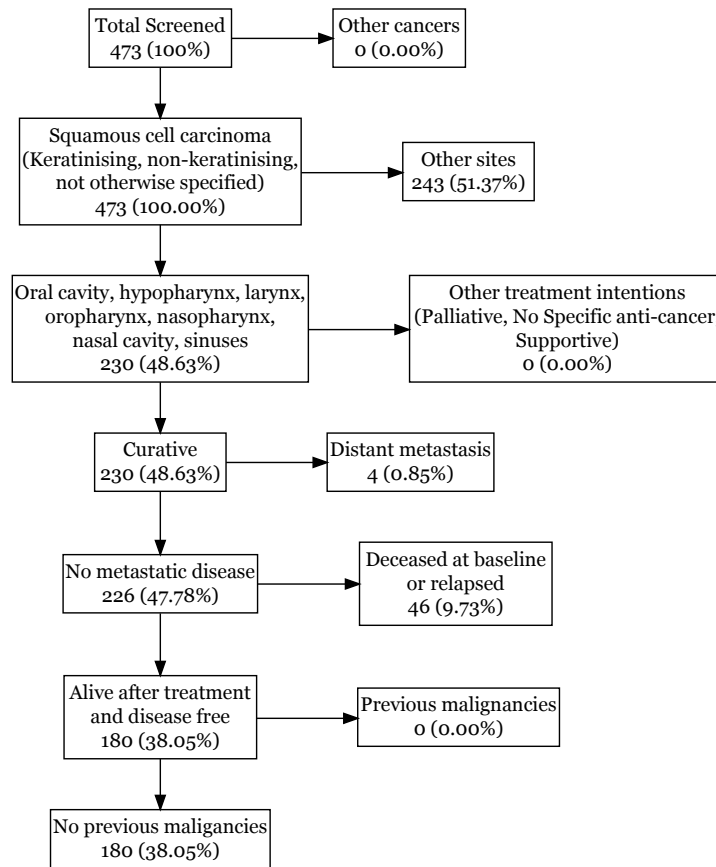
77 **Figure 3 - EORTC QLQ-H&N35 scales histograms.**



78

79

Figure S4 – INT cohort inclusion flowchart.

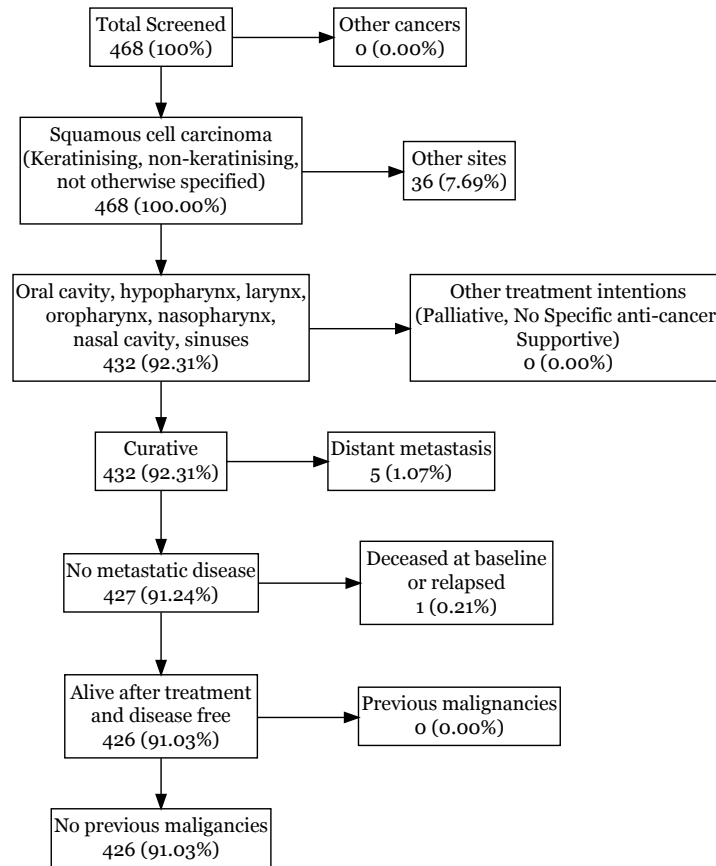


80

81

82

Figure S5 – UMM1 cohort flowchart.



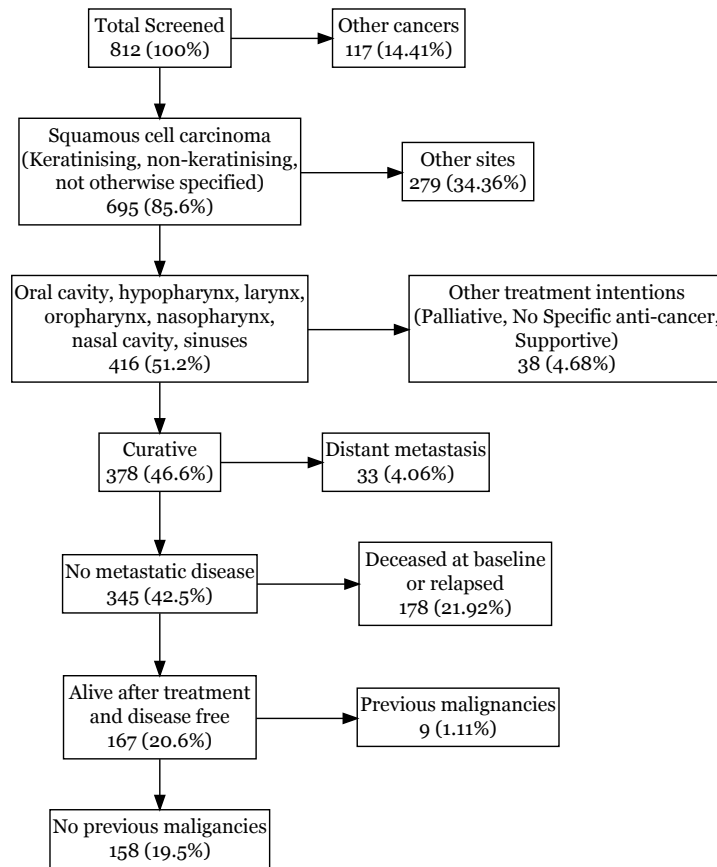
83

84

85

86

Figure S6 - UMM2 cohort flowchart.

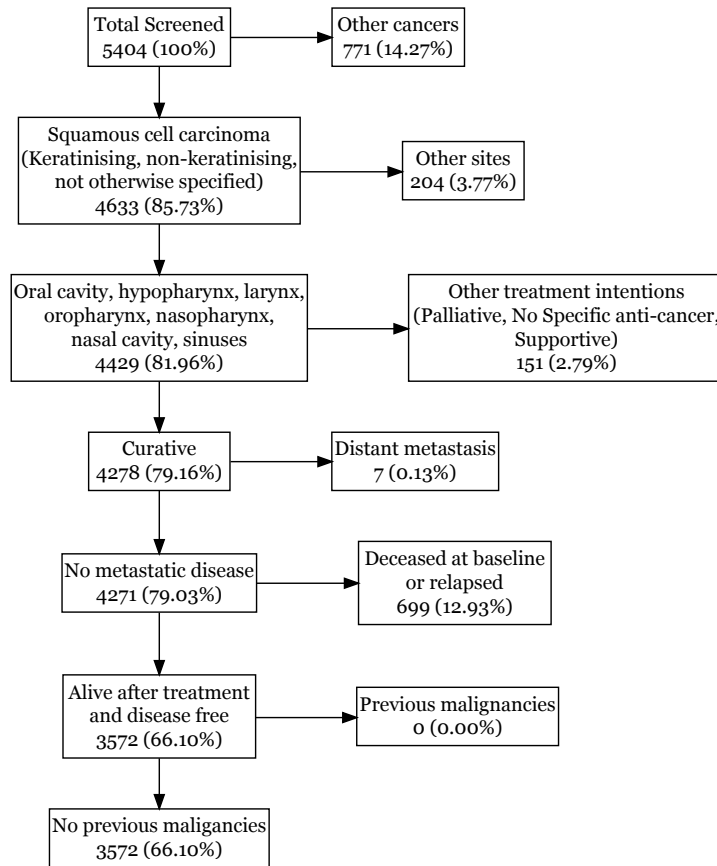


87

88

89

Figure S7 - UMM3 cohort flowchart.



90

91

Figure 8 - HN5000 cohort flowchart.