SHORT COMMUNICATION

Use of an Artificial Neural Network to Identify Patient Clusters in a Large Cohort of Patients with Melanoma by Simultaneous Analysis of Costs and Clinical Characteristics

Giovanni DAMIANI[1–3], Alessandra BUJA[4]*, Enzo GROSSI[5], Michele RIVERA[4], Anna DE POLO[4], Giuseppe DE LUCA[4], Manuel ZORZI[6], Antonella VECCHIATO[7], Paolo DEL FIORE[7], Mario SAIA[8], Vincenzo BALDO[4], Massimo RUGGE[6], Carlo Riccardo ROSSI[7] and Gianfranco DAMIANI[9,10]

*[1]Clinical Dermatology, IRCCS Istituto Ortopedico Galeazzi, [2]Department of Biomedical, Surgical and Dental Sciences, University of Milan, Milan, [3]PhD Degree Program in Pharmacological Sciences, Department of Pharmaceutical and Pharmacological Sciences, University of Padua, Padua, [4]Department of Cardiological, Vascular, and Thoracic Sciences, and Public Health, University of Padua, Via Loredan 18, IT-35128 Padova, [5]Villa Santa Maria Institute, Neuropsychiatric Rehabilitation Center, Tavernerio (Como), [6]Veneto Tumor Registry, Azienda Zero, [7]Surgical Oncology Unit, Veneto Institute of Oncology IOV-IRCCS, [8]Veneto Regional Authority, Padua, [9]Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, and [10]Università Cattolica del Sacro Cuore, Rome, Italy. *E-mail: alessandra.buja@unipd.it*

Accepted Oct 28, 2020; Epub ahead of print Nov 2, 2020

The incidence of cutaneous malignant melanoma (CMM) in Italy has increased in the last decade, leading to public health concern and rising costs of healthcare (1, 2). In addition to individual susceptibility to development of CMM, several environmental variables influence prognosis in this disease. These variables include social disparities, socioeconomic status, education and marital status (3). However, the impact of these variables on costs is unknown. The current study used a new methodology, based on an artificial neural network (ANN), to decodify this complexity by simultaneously describing the relationships between clinical, sociodemographic, outcome, and cost variables, and grouping patients into clusters (4, 5).

MATERIALS, METHODS AND RESULTS

This study evaluated a collaborative registry of 556 patients (Veneto Tumor Registry & Veneto Oncology Network)[1], who were diagnosed with CMM by a board certified dermatopathologist in 2015 in 4 of the 7 provinces of the Veneto Region in Northern Italy (3). For each patient, the CMM registry includes a set of tumour characteristics, including: tumour-node-metastasis (TNM) stage at diagnosis; Breslow thickness (mm); Clark's level of invasion (I–V); presence of ulceration (yes/no); site (trunk, head, limbs); cost categories tertiles (scintigraphic, surgical, medical, instrumental, cyto/histological, microbiological, blood examinations, radiotherapeutic, radiological and total (costs based on "Hospital Discharge Forms" (SDO)) CMM-specific mortality.

Costs were assessed from the perspective of the Italian National Health System (Italian NHS), taking only direct costs into account. Patients were linked via unique anonymous identification codes to all administrative data regarding their hospital admissions, day hospital service use, drug usage, visits to emergency services, medical devices used at home, and hospice admissions. These data were used to compute the direct costs for each patient in the 5 years after diagnoses of their CMM.

Descriptive analyses were performed using absolute and relative frequencies for categorical variables. A semantic connectivity map was constructed using Auto Contractive Map (AutoCM, Semeion©, Rome, Italy) to elucidate variable links (4). The system highlights the natural links on a graph based and distances between variables reflect the weights of the ANN (Appendix S1[2]). AutoCM has many relevant features: (*i*) non-linear associations among variables are preserved; (*ii*) patterns of connections between clusters of variables are captured; and (*iii*) complex similarities among variables emerge.

Clinical/histological and demographic data for the 556 included CMM are summarized in Table SI[2].

The AutoCM results are shown in **Fig. 1**. The item "1–6 mitoses" is the centre (main attractor) of our unsupervised analysis, demonstrating its clustering value *vis-à-vis* the spread of its 4 main branches (strength >0.60) depicting 4 endotypes. The use of radiotherapy, education and marital status were central descriptors in our database. The 1st endotype grouped together (or "clusters") those patients with advance-stage CMM who had nodular and ulcerated CMM, a high risk of death, and a heavy economic burden. The 2nd endotype clustered patients >60 years of age who had CMM on the trunk or face, and high procedural and therapeutic costs. The 3rd endotype clustered patients with stage Ib CMM. The 4th endotype clustered patients with no radiotherapy costs, comprising 4 main subsets, each with their own biological and socioeconomic variable items.

Further details of the items interactions are shown in Appendix S2[2]. Our results, in-line with the idea of precision medicine, also suggest a potential endotype-guided treatment Appendix S3[2] that may implement CMM follow-up visits.

DISCUSSION

Through a combined analysis of the clinical, sociodemographic and economic variables associated with CMM, using an ANN, endotypes were identified that can be used to estimate an individual patient's final costs based on their baseline characteristics. Based on these patient clusters, further tests can be suggested to add to the routine tests used for patients with CMM at each TNM stage.

Use of a machine learning approach such as this can generate a comprehensive model that establishes complex links between clinical, sociodemographic, outcome, and cost variables. Machine learning has also been used by Finlayson et al. (6) on a small database containing the clinical, therapeutic and molecular features of 237 cases of advanced CMM, with the aim of helping clinicians predict patients survival and make therapeutic decisions.

In our database the main variable, also regarded as the "attractor", was "1–6 mitoses. Mitotic index is not included in the 8th edition of the American Joint Committee on Cancer TNM system, suggesting that it is optional to report this variable for prognostic purposes. The value of the mitotic index as a prognostic indicator in CMM has been debated, partly because of its moderate interobserver

[1]The dataset generated during the current study is not publicly available but is available from the corresponding author (alessandra.buja@unipd.it) on reasonable request.

[2]https://www.medicaljournals.se/acta/content/abstract/10.2340/00015555-3680

doi: 10.2340/00015555-3680
Acta Derm Venereol 2020; 100: advxxxxx

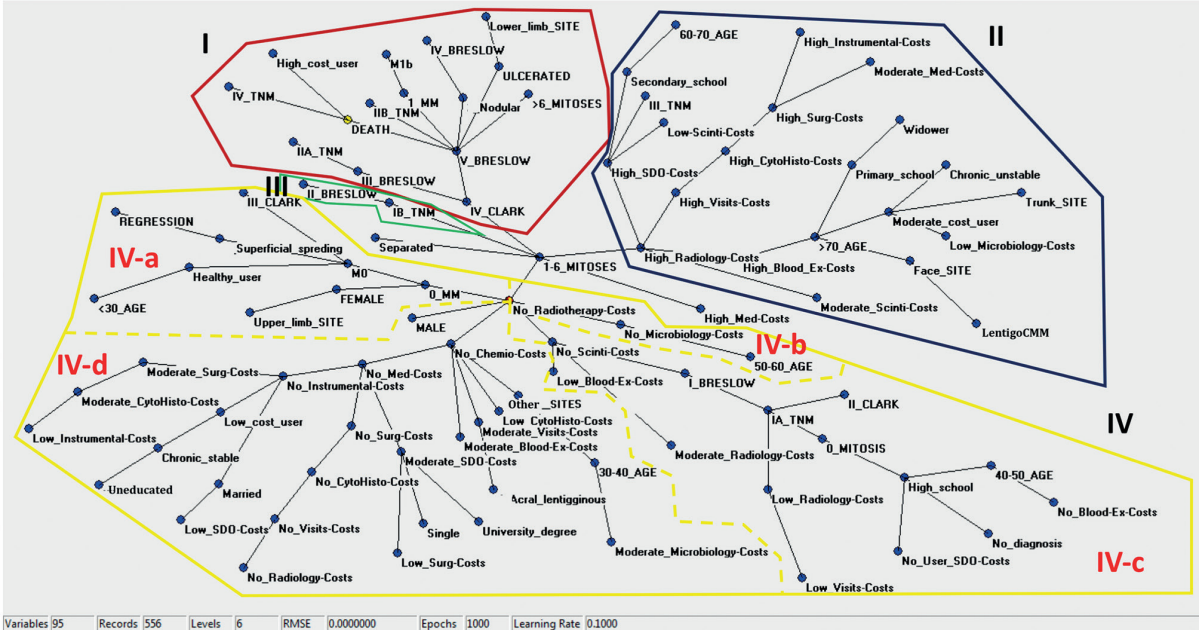

Fig. 1. Auto Contractive Map (AutoCM) semantic map, showing cutaneous malignant melanoma (CMM) endotypes created by clustering variables belonging to different fields (clinical, therapeutic, histological, demographic and costs). "0_MM": absence of metastases; "1_MM": metastases clinically and/or radiologically evident; Breslow_1: <0.76 mm; "Breslow_2": 0.76–1.75 mm; "Breslow_3": >1.75 mm; "Breslow_4": missing data; "Breslow_5": regression or metastases; "CMM": cutaneous malignant melanoma; "Death": melanoma-specific mortality; No diagnosis: no comorbidities. Note: it was opted to exclude branches radiating from "1–6 mitoses" (which has a bond strength <0.60) as endotypes, so the branches "separated" and "high costs of medical therapies" were excluded. Four main endotypes were identified (I, II, III, IV). The fourth endotype has 4 sub-endotypes, which were termed IV-a, IV-b, IV-c, IV-d. The endotypes are as follows: (I) advanced-stage patients: with nodular and ulcerated CMM on lower limbs (Clark IV, Breslow III–IV, TNM IV or II with metastases and >6 mitoses), high risk of death, and heavy economic burden; (II) patients >60 years, widowed, with primary school diploma, with unstable chronic disease, non-nodular CMM on the trunk, or lentigo maligna on the face, with 1–6 mitoses, associated with high costs of SDO (Hospital Discharge Form), diagnostic procedures (visits, cytohistology, radiological and instrumental investigations), and therapies (medical, surgical); (III) patients with stage Ib disease: patients with TNM Ib, and Breslow II with 1–6 mitoses; (IV) a cluster around the absence of costs for radiotherapy, which includes 4 main subsets, each with their own biological and socioeconomic items: (IV-a) female under 30 years old with non-metastatic, regressive and superficial spreading CMM of the upper limb, with 1–6 mitoses; (IV-b) male aged 50–60 years with CMM with 1–6 mitoses, and without microbiological costs; (IV-c) male aged 40–50 years with a high-school diploma, with TNM Ia, Clark II, Breslow I CMM, with no mitosis, associated with moderate radiology costs; (IV-d) a miscellany of CMM patients with 1–6 mitoses, further classifiable as: (a) married and uneducated, with a stable chronic condition, which was associated with moderate costs of cytohistology, surgery, low costs of instrumental investigations, and low total costs; (b) single with a university degree, associated with moderate hospital discharge records (SDO), and low costs of surgery; (c) male aged 30–40 years with acral lentiginous CMM, with 1–6 mitoses, associated with moderate costs of blood tests, specialist visits, and microbiological tests.

variability (7). The findings of the current study suggest that the mitotic index is relevant for predicting the costs of CMM from baseline information.

The semantic map used in the current study revealed 4 endotypes, through using the database as "learning material" for the evolutive algorithm of the ANN (8). The internal validity of the algorithm is high and measurable; however, further study is required to determine its external validity, since the ANN was designed to find connections among variables in the database in the current study. Evidence to support external validity come from the literature, specifically regarding marital status and education. In line with literature (3), we confirmed that marital status represents a valuable information to be recorded in CMM patients. An association between married status and lower costs may stem from the presence of a partner resulting in earlier diagnosis of a skin cancer. This suggests that patients with little or no formal education are less inclined to access healthcare services (3).

This study has some limitations. Data regarding the patients' self-reported educational level may not be reliable, although a previous Italian study on the validity of data on education levels recorded in hospital discharge records found that it was in the good-to-excellent range (9). The current study lacks information on patients' lifestyles and other socioeconomic parameters that might reduce the influence of educational level on the natural history of their CMM.

A strength of this study is its population-based dimension, which minimizes selection bias by using independently-acquired administrative data. The socioeconomic impact of education level and/or income is likely to be mitigated in a universalistic health system like the Italian NHS.

In conclusion, in CMM, clinical variables together with costs were indispensable to cluster patients in endotypes by ANN. Endotypes-guided management is affirming as new promising strategy to guide medical and surgical therapies (10, 11).

ACKNOWLEDGEMENTS

IRB approval. Veneto Oncological Institute Ethics Committee (n° 695/20.10.2016).

The authors have no conflicts of interest to declare.

REFERENCES

1. Global Burden of Disease Cancer Collaboration, Fitzmaurice C, Abate D, Abbasi N, Abbastabar H, Abd-Allah F, et al. Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2017: a systematic analysis for the Global Burden of Disease Study. JAMA Oncol 2019; 5: 1749–1768.
2. GBD 2017 Italy Collaborators. Italy's health performance, 1990–2017: findings from the Global Burden of Disease Study 2017. Lancet Public Health 2019; 4: e645–e657.
3. Buja A, Lago L, Lago S, Vinelli A, Zanardo C, Baldo V. Marital status and stage of cancer at diagnosis: a systematic review. Eur J Cancer Care (Engl) 2018; 27: doi: 10.1111/ecc.12755.
4. SEMEION, AUTOCM – AutoCM – Auto Contractive Map (Semeion©). [accessed 2020 March 6]. Available from: http://www.semeion.it/wordpress/en/AutoCM/
5. Damiani G, Grossi E, Berti E, Conic RRZ, Radhakrishna U, Pacifico A, et al. Artificial neural networks allow response prediction in squamous cell carcinoma of the scalp treated with radiotherapy. J Eur Acad Dermatol Venereol 2020; 34: 1369–1373.
6. Finlayson SG, Levy M, Reddy S, Rubin DL. Toward rapid learning in cancer treatment selection: an analytical engine for practice-based clinical data. J Biomed Inform 2016; 60: 104–113.
7. Saldanha G, Ali R, Bakshi A, Basiouni A, Bishop R, Colloby P, et al. Global and mitosis-specific interobserver variation in mitotic count scoring and implications for malignant melanoma staging. Histopathology 2020; 76: 803–813.
8. Grossi E. Artificial adaptive systems and predictive medicine: a revolutionary paradigm shift. Immun Ageing 2010; Suppl 1: S3.
9. Ventura M, Colais P, Fusco D, Agabiti N, Cesaroni G, Davoli M. Information on educational level from hospital discharge register: an analysis of validity. Epidemiol Prev 2013; 37: 289–296.
10. Grossi E, Stoccoro A, Tannorella P, Migliore L, Coppedè F. Artificial neural networks link one-carbon metabolism to gene-promoter methylation in Alzheimer's disease. J Alzheimers Dis 2016; 53: 1517–1522.
11. Bartoloni E, Baldini C, Ferro F, Alunno A, Carubbi F, Cafaro G, et al. Application of artificial neural network analysis in the evaluation of cardiovascular risk in primary Sjögren's syndrome: a novel pathogenetic scenario? Clin Exp Rheumatol 2019; 37: 133–139.

Appendix S1.

SUPPLEMENTARY MATERIAL AND METHODS

Auto Contractive Map insights

The Auto Contractive Map (AutoCM, Semeion©, Rome, Italy) is a fourth-generation unsupervised artificial neural network (ANN) that can be used to assess databases, elucidating the network of links between considered variables and enabling network-based clustering of patients.

The database is regarded as a series of matrixes and weighted, it is filtered using a minimum spanning tree algorithm (MST), generating a graph that contains biological evidence which has already been tested in the medical field (S1–S4). The aim of this data-mining model is to reveal hidden trends and associations among variables. The algorithm can be used to create a semantic connectivity map in which non-linear associations are preserved and explicit connection patterns are described. This approach maps relevant connections between and among variables, and the principal hubs of the system. Hubs can be defined as variables with the maximum number of connections in the map. From a mathematical standpoint, the specificity of the AutoCM algorithm lies in its ability to minimize a complex cost function by comparison with traditional algorithms.

Traditional cost minimization function:

$$E = Min\left\{\sum_{i}^{N}\sum_{j}^{N}\sum_{q}^{M} u_i^q \cdot u_j^q \cdot \sigma_{i,j}\right\}$$

AutoCM cost minimization function:

$$E = Min\left\{\sum_{i}^{N}\sum_{j}^{N}\sum_{k}^{N}\sum_{q}^{M} u_i^q \cdot u_j^q \cdot u_k^q \cdot A_{i,j} \cdot A_{i,k}\right\};$$

$$\mathbf{A} = \left(1.0 - \frac{\mathbf{w}}{C}\right);$$

N = Number of Variables (Columns);

M = Number of Patterns (Rows).

Comparing the 2 cost functions, traditional minimization includes only second-order effects, whereas the AutoCM algorithm also considers a third-order effect. Thus, the AutoCM algorithm can reveal similarities between variables that are completely embedded in the dataset and invisible to other, classical, tools. This approach describes a context typical of living systems, in which there is continuous, time-dependent complex change in the value of the variables. AutoCM can also learn under difficult circumstances, when the connections on the main diagonal of the second connections matrix are removed, for instance. When the learning process is organized in this way, AutoCM identifies specific relationships between each variable and all the others. From an experimental point of view, the ranking of its connections matrix consequently seems to equate to the ranking of the joint probability between each variable and the others. AutoCM requires a training phase in which the algorithm learns how variables are interconnected. The AutoCM's learning algorithm can be summarized in 4 orderly steps: (a) signal transfer from the input into the hidden layer; (b) adaptation of the connections value between the input layer and the hidden layer; (c) signal transfer from the hidden layer into the output layer; (d) adaptation of the connections value between the hidden layer and the output layer.

The MST equates to the "nervous system" of a given dataset. From the sum of all the connection strengths among all the variables, the total energy of the system is obtained. The MST selects only the connections that minimize this energy, which are the only ones needed to keep the system coherent. So, all the links included in the MST are fundamental, but not every "fundamental" link in the dataset needs to be in the MST. This limitation is intrinsic in the nature of the MST: every link that gives rise to a cycle in the graph (viz., that destroys the graph's tree-like shape) is removed, whatever its strength and meaningfulness. To fix this shortcoming, and better capture the intrinsic complexity of a dataset, more links need to be added to the MST, based on 2 criteria: (*i*) the new links have to be relevant in the quantitative sense; and (*ii*) from the qualitative standpoint, they have to be able to generate new, regular cyclic microstructures. The additional links superimposed on the MST graph generate a maximally regular graph (MRG).

This MRG is the graph with the hub function achieving the highest value among all the graphs generated by putting the connections previously skipped during the computation of the MST back into the original MST, one by one. In other words, starting from the MST, the MRG presents the largest number of regular microstructures, highlighting the most important connections of the dataset. The resulting "diamond" expresses the core complexity of the system and, in our particular case, the core of the disease.

REFERENCES

S1. Buscema M, Grossi E. The semantic connectivity map: an adapting self-organising knowledge discovery method in data bases. Experience in gastro-oesophageal reflux disease. Int J Data Min Bioinform 2008; 2: 362–404.

S2. Buscema M, Grossi E, Snowdon D, Antuono P. Auto-Contractive Maps: an artificial adaptive system for data mining. An application to Alzheimer disease. Curr Alzheimer Res 2008; 5: 481–498.

S3. Buscema M. A novel adapting mapping method for emergent properties discovery in data bases: experience in medical field. Proceeding of IEEE International Conference on Systems, Man and Cybernetics (SMC 2007)", October 2007. IEEE, Montreal, Canada 2007: 3457–3463.

S4. Licastro F, Porcellini E, Chiappelli M, Forti P, Buscema M, Ravaglia G, Grossi E. Multivariable network associated with cognitive decline and dementia. Neurobiol Aging 2010; 31: 257–269.

Appendix S2.

SUPPLEMENTARY RESULTS

Detailed evaluation of the AutoCM analysis

It is notable that the so-called "diamond" describing the higher rank of interrelationships between items developed entirely within endotype IV, with both its economic and clinical centres in "no costs of radiotherapy", "no costs of instrumental investigations", "no costs of medical therapies", "no costs of chemotherapies", "no costs of scintigraphy", "no costs of microbiological tests" and "no metastases", "no mitoses", "TNM Ia", "Breslow I", "non-specific site", and "superficial spreading".

Appendix S3.

SUPPLEMENTARY RESULTS

Endotypes clinical implications

The current study reports only preliminary results; however, the authors propose some potential clinical implications of the endotypes formed.

Endotype I: Once a primary lesion has metastasized, clinical and dermatoscopic monitoring of new skin and mucosal lesions is recommended, together with testing for occult blood in the stool.

Endotype II: For patients with unstable, chronic disease, dermatologists and oncologists should interact more closely with other specialists and the family doctor.

Endotype IV-c: Given the moderate costs, it is recommended to add the following to the routine tests: chest X-ray; urine microbiological testing; C-reactive protein (CRP) and procalcitonin (if CRP levels are altered) to rule out bacterial infections.

Endotypes III, IV-a, IV-b, IV-d: Wait and see.

These suggestions also need further validation in other cohorts with the same clinical/economic and legislative context.

Table SI. Clinical/histological and demographic characteristics of the enrolled patients with cutaneous malignant melanoma (CMM) who presented on the index date (CMM date)

Characteristics	%	*n*
Sex		
Male	47.66	265
Female	52.34	291
Age		
Under 30 years	3.78	21
30–40 years	9.53	53
40–50 years	24.64	137
50–60 years	20.68	115
60–70 years	17.99	100
Over 70 years	23.38	130
Married status		
Single	7.37	41
Married	30.22	168
Separated	2.16	12
Widowed	1.98	11
Undefined	58.27	324
Education		
Uneducated	7.91	44
Primary school	16.01	89
Secondary school	24.46	136
High school	41.91	238
University degree	8.81	49
T stage		
IA	58.63	326
IB	24.10	134
IIA	5.22	29
IIB	3.96	22
III	6.83	38
IV	1.26	7
Metastases		
Yes	3.80	21
No	96.20	535
Chronic condition[a]		
No chronic conditions	79.32	441
Chronic, stable	12.95	72
Chronic, unstable	7.73	43
Site (label assigned in AutoCM)		
Back (trunk)	29.86	166
Upper limbs (limbs)	18.53	103
Thorax (trunk)	8.45	47
Face (head)	6.66	37
Anterior thigh (limbs)	6.66	37
Armpits (limbs)	5.40	30
Medial leg (limbs)	5.22	29
Posterior leg (limbs)	4.68	26
Abdomen (other sites)	4.32	24
Posterior thigh (limbs)	2.52	14
Ears (head)	1.80	10
Instep (other sites)	1.80	10
Sole (other sites)	1.80	10
Scalp (head)	1.26	7
Gluteal area (other sites)	0.90	5
Anus (other sites)	0.18	1

Table SI. contd.

Characteristics	%	*n*
Histotype		
Superficial spreading	71.94	400
Nodular	7.19	40
Lentigo maligna melanoma	2.16	12
Acral lentiginous	2.16	12
Not specified	16.55	101
Clark's level of invasion		
I	2.88	16
II	36.51	203
III	39.75	221
IV	20.86	116
Ulceration		
Ulcerated	11.51	64
Not ulcerated	88.49	492
Mitoses		
0	60.61	337
1–6	33.09	184
>6	6.29	35
Regression		
No regression	70.50	392
Regression	29.50	164
Breslow thickness[b]		
Mean ± SD, mm	1.12 ± 1.90	
I (<0.76 mm)	65.47	364
II (0.76–1.75 mm)	17.81	99
III (>1.75 mm)	5.04	28
IV (missing data)	2.70	15
V (regression or metastases)	8.99	50

[a]Chronic conditions were retrieved using ICD10-codes and the Chronic Condition Indicator for the International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) (https://www.hcup-us.ahrq.gov/toolssoftware/chronic_icd10/chronic_icd10.jsp#overview). In the Hospital Discharge Form (SDO) codes used in the administrative database for hospitalization, there is a different code for stable and unstable disease, hence the criteria to define stability or instability are clinical ones and in accordance with the adopted disease-specific guidelines. [b]These Breslow categories are coded in the Italian Nomenclature for Pathology Codes.
SD: standard deviation; AutoCM: Auto Contractive Map (AutoCM, Semeion©, Rome, Italy).