Cross-Modal Retrieval of Urban Scene Characteristics Using Natural Language Descriptions and Street View Images

Eivor Strom
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
helloeivor@unr.edu

Abstract

Natural language provides an expressive interface for searching large street view image archives, yet most computational studies of urban scenes reduce visual content to fixed scores such as greenery, enclosure, or perceived safety. Such scores are useful for audit-style measurement but do not support open-ended retrieval by combinations of visible characteristics. This paper develops a cross-modal retrieval framework that maps free-text descriptions of urban scene characteristics to street view images. The framework combines a scene-characteristic ontology with a dual-encoder retrieval model, attribute-aware contrastive learning, hard negative construction, and uncertainty-aware ranking. Queries are decomposed into semantic attributes such as vegetation, facade complexity, sky openness, street furniture, maintenance, human activity, and walkability cues; image and text encoders are trained to place compatible descriptions and views in a shared embedding space. An attribute-conditioned similarity score preserves interpretable control over retrieval, while hard negatives reduce reliance on superficial lexical or geographic shortcuts. A calibration layer separates similarity from confidence and supports abstention when scene evidence is weak or viewpoint-dependent. The paper specifies the retrieval objective, loss terms, ranking logic, evaluation metrics, and principal failure modes, including ontology mismatch, false negatives in many-to-many urban settings, and spatial leakage during validation. The contribution is a technically grounded retrieval design for planning, public health, and urban analytics that treats natural language as a first-class query modality rather than as auxiliary metadata.

Keywords

cross-modal retrieval; street view imagery; natural language; urban scene characteristics; contrastive learning; uncertainty calibration

1. Introduction

Street view imagery has become a routine data source for observing urban environments at scales that manual field audits cannot match [1], [2]. Image archives now cover many cities across multiple years, and computer vision pipelines can extract visible features such as buildings, roads, vegetation, sky, and pedestrians. These capabilities have supported studies of neighborhood demographics, physical disorder, perceived safety, and walkability. The central analytical opportunity is not merely to classify pixels but to answer questions about the visible character of places: where mature tree canopies enclose narrow residential streets, where continuous storefronts and street-level activity coexist, or where large setbacks and blank facades produce low pedestrian support.

Computational work on urban perception has demonstrated that deep networks can predict human ratings and physical-environment indicators from street-level imagery [3], [4], [5]. However, fixed-target prediction and retrieval are different tasks. A model trained to score greenery or enclosure can answer only pre-specified questions, and its output is a scalar or a label attached to an image. A planner, researcher, or community group may instead ask for images that match a composite description, compare characteristic combinations across neighborhoods, or search for rare configurations that were never defined as separate training labels. Natural language is attractive in this setting because it can express relations, intensities, co-occurrence constraints, and context-dependent expectations without requiring a new classifier for every query.

Urban design theory supplies part of the conceptual vocabulary for such queries. Legibility, imageability, enclosure, human scale, transparency, complexity, and coherence describe experiential qualities that are visible in street-level views but are not identical to object categories [6], [7]. These qualities are relational: a wide street may feel enclosed if buildings are tall and continuous, while the same building height may feel exposed if setbacks are large and frontages are interrupted. Retrieval should therefore preserve both attribute-level evidence and the composition of attributes in a scene. A system that represents only global scene similarity will struggle when users care about why an image is relevant.

The contribution developed here is an attribute-aware cross-modal retrieval framework for natural language descriptions and street view images. It formulates retrieval as ranking over an image collection conditioned on a textual scene description; represents descriptions through a scene-characteristic ontology; learns a shared image-text embedding with contrastive and attribute-alignment terms; constructs hard negatives that target compositional and geographic shortcuts; and calibrates similarity scores so that relevance and confidence are separated. The manuscript is methodological and analytical. It does not report a deployed system or benchmark results; rather, it specifies a technically coherent design, derives the consequences of its modeling choices, and identifies evaluation logic and boundary conditions that empirical work would need to satisfy.

2. Related Work and Problem Formulation

2.1 Street view imagery and scene characteristics

Street view imagery supports scalable observation of visible urban form, but its use requires attention to sampling, viewpoint, and construct validity [2], [3]. Reviews of street view applications show recurring interest in greenery, walkability, disorder, safety, and aesthetic quality, often measured by human audits or supervised models [2], [3]. Deep-learning studies have extended this line by predicting perceptual ratings and demographic associations from image features [1], [4], [5]. These studies establish that street-level visual signals carry information about urban conditions, yet they generally assume a closed set of target variables. The retrieval problem instead requires a representation that can be queried with new combinations of descriptors and can return ranked visual evidence.

Scene characteristics should not be treated as pure object counts. Enclosure depends on the ratio of vertical surfaces to street width and sky exposure; facade complexity depends on the number, scale, and contrast of architectural elements; maintenance depends on surface condition, vegetation management, and visible repair; walkability cues include continuity of frontage, sidewalk presence, street furniture, and indicators of use. Some characteristics are directly observable, while others are inferred from visual cues and are therefore uncertain. A

retrieval framework must preserve this distinction, because a query about perceived safety or social atmosphere is not equivalent to a query about the presence of a bench or a tree.

2.2 Cross-modal retrieval and the retrieval objective

Vision-language representation learning has shown that large image-text collections can train encoders whose embeddings support zero-shot recognition and retrieval [8]. Dense retrieval methods similarly demonstrate that learned embeddings can support semantic search over large collections [9]. Urban applications have begun to combine street view imagery with other modalities: one recent study uses geotagged social media posts and street view imagery with deep neural networks to model urban streetscapes [10]. That work is relevant because it treats visual and textual urban signals jointly, but it is not an open-vocabulary text-to-image retrieval formulation; the textual signal is used for modeling rather than for ranking street view images against arbitrary natural language queries.

Formally, let I be an indexed collection of street view images, let x_i denote the image with index i, and let q be a natural language description. A retrieval system returns a ranked subset of images according to a similarity function $s(q, x_i)$. The basic retrieval rule is:

$$R(q) = argmax_i \in I\ s(q, x_i)$$

Here R(q) denotes the image or ranked list selected for query q, and the argmax is taken over the candidate image collection I. In practice, retrieval returns the top-k images rather than a single winner, because many street segments may partially match a composite description. The ranking assumption is that higher $s(q, x_i)$ indicates stronger correspondence between the described scene characteristics and the visible content of the image. The central modeling difficulty is that q and x_i inhabit different spaces: text describes characteristics and relations, while images contain pixels, viewpoint effects, occlusions, and incidental objects.

3. Methodological Framework

3.1 Scene characteristic ontology and query representation

The framework begins with a controlled but extensible ontology of K scene characteristics. Each characteristic k has a name, a definition, a set of positive visual cues, a set of exclusion conditions, and a scale indicating whether it is directly visible or inferential. Greenery, for example, may include trees, grass, and planters, but its intensity depends on canopy coverage and vertical layering. Enclosure may depend on building height, street width, setback continuity, and sky visibility. Human activity may include pedestrians, seating, and open storefronts, but its interpretation depends on time of day and image timing.

Queries are parsed into a global descriptive vector and K attribute-specific mentions. A query such as “a narrow street with mature trees, low traffic, and active ground-floor frontages” contains a global scene type, an enclosure cue, a vegetation cue, a traffic cue, and an activity cue. The text encoder receives the full description and optional structured attribute tokens. This dual treatment allows the model to learn holistic phrasing while retaining attribute-level control. The ontology also supports negative constraints, such as “not a wide arterial” or “no construction barriers,” which are important because retrieval users often define a scene by what it is not.

3.2 Dual-encoder similarity and indexing

The retrieval model uses a text encoder $g\theta$ and an image encoder $h\varphi$. The text encoder maps a query q to a normalized embedding, and the image encoder maps a street view image x_i to a normalized embedding in the same space. Similarity is computed by cosine distance:

$$s(q, x_i) = \frac{\left(g\theta(q) \cdot h\varphi(x_i)\right)}{(||g\theta(q)|| ||h\varphi(x_i)||)}$$

In this expression, $g\theta(q)$ is the query embedding produced by the text encoder with parameters θ, $h\varphi(x_i)$ is the image embedding produced by the image encoder with parameters φ, the centered dot denotes an inner product, and $\|\cdot\|$ denotes the Euclidean norm. The normalization constrains similarity to a bounded range and stabilizes contrastive training. The dual-encoder design permits all image embeddings to be computed offline and indexed, so a new textual query requires only one text-encoder pass and a nearest-neighbor search. This computational property is important for city-scale archives, where exhaustive image-text scoring at query time would be prohibitive.

The image encoder may operate on a full street view or on a set of perspective crops. Full-image encoding preserves composition, while crop-level encoding can represent sidewalk, facade, and sky regions separately. A practical design aggregates crop embeddings with attention weights conditioned on the query attributes, but the global vector remains the primary index for efficiency. This trade-off follows established two-branch retrieval practice, where separate encoders are trained to preserve semantic correspondence without requiring a joint attention pass over every candidate at inference time [8], [9].

3.3 Attribute-aware contrastive learning

The primary training signal is a contrastive objective over matched query-image pairs. For a batch of B pairs, the query-to-image loss encourages each query to be closer to its matched image than to other images in the batch:

$$L_ret = -(1/B) \sum_b \log\left(\exp(s(q_b, x_b)/\tau) \,/\, \sum_j \exp(s(q_b, x_j)/\tau) \right)$$

Here q_b is the query in pair b, x_b is its matched image, j ranges over the batch candidates, and τ is a temperature parameter that controls the sharpness of the distribution. A symmetric image-to-query term can be added in the same manner. The batch composition matters: if negatives are sampled only from unrelated cities or radically different scene types, the model may learn coarse distinctions while remaining insensitive to fine-grained attribute composition. Hard negative construction therefore supplements the basic contrastive loss.

Contrastive objectives of this form are standard in cross-modal embedding work [11]. Triplet margins provide an additional way to separate positive and negative pairs [12], and large-batch visual contrastive learning shows how negative sampling shapes representation quality [13]. Two-branch image-text matching architectures further support the idea that separate modality-specific projections can be aligned through a shared ranking objective [14]. The urban setting adds an attribute constraint. The model should not treat two images as interchangeable merely because both contain trees or both contain pedestrians; the query may depend on the relationship among vegetation, street width, frontage continuity, and activity.

An attribute-aware similarity combines the global score with attribute-specific scores:

$$f(q, x_i) = \lambda s(q, x_i) + (1 - \lambda) \sum_k \alpha_k s_k(q, x_i)$$

In this expression, $f(q, x_i)$ is the composite retrieval score, $s_k(q, x_i)$ is the similarity for attribute k, α_k is a nonnegative attribute weight, and $\lambda \in [0,1]$ balances holistic and attribute-level evidence. The weights can be set by query parsing, by user controls, or by a validation procedure. This formulation provides a transparent mechanism for expressing that a query about enclosure should not be satisfied by a wide street with tall isolated towers, even if both contain tall buildings. It also allows an analyst to test sensitivity to attribute weighting rather than relying on a single opaque ranking.

3.4 Hard negative construction and margin

Hard negatives are candidate images that are plausible under superficial cues but incorrect under the full query. In an urban archive, hard negatives may share the same neighborhood, camera direction, or dominant object with the positive image while differing in the characteristic combination that matters. Examples include an image with trees but no street enclosure, an image with storefronts but heavy traffic barriers, or an image with a sidewalk but no pedestrian-scale frontage. A margin loss can explicitly penalize such confusions:

$$L_H N = \max\big(0, \mu + s(q, x^-) - s(q, x^+)\big)$$

Here x^+ is the positive image for query q, x^- is a hard negative, and μ is a margin that requires the positive score to exceed the negative score by at least μ. The loss is zero when the margin is satisfied. Hard negatives should be generated with care because urban descriptions are many-to-many: two different street segments may both satisfy "quiet residential street with mature trees," even if only one is labeled positive. Overly aggressive negative mining can therefore teach the model to reject valid alternatives. A practical safeguard is to mark high-similarity candidates as ambiguous unless they violate explicit negative constraints in the query.

Hard-negative selection also interacts with textual supervision quality. Large-scale vision-language pre-training indicates that noisy or loosely matched text can still support useful representations when training and filtering are designed appropriately [15]. In the urban case, descriptions may come from audits, planning documents, or user queries, each with different levels of specificity. The loss should therefore allow soft targets for partially relevant images rather than treating every non-positive image as equally negative.

3.5 Uncertainty calibration and retrieval decision

A high similarity score does not guarantee that the image actually exhibits the requested characteristic. Occlusion, season, camera angle, and ontology ambiguity can all produce confident but incorrect rankings. Calibration separates ranking from confidence. Let $u(q, x_i)$ denote an uncertainty score for the pair. A logistic calibration layer can transform the raw similarity into a relevance probability:

$$u(q, x_i) = 1 - \sigma(as(q, x_i) + b)$$

In this expression, σ is the logistic function, and a and b are calibration parameters that would be estimated on held-out query-image pairs with relevance judgments. Before such estimation, they are design parameters, not empirical claims. Modern neural networks are often miscalibrated, so a separate calibration stage is justified when retrieval outputs inform decisions [16]. Contrastive representation learning also provides a general account of discrimination under noisy correspondence, which supports treating similarity as a learned ranking signal rather than a directly interpretable probability [17].

Calibration enables a retrieval policy that exposes only images satisfying both a relevance and a confidence condition. For example, an image may be returned when its composite score exceeds a query-specific threshold and its uncertainty remains below a tolerance. Images that fail the confidence condition can be placed in a review queue or omitted from an automated summary. This is especially important for queries involving inferential characteristics such as perceived safety, where visual evidence may be insufficient. The policy does not eliminate error, but it makes the trade-off between coverage and precision explicit.

3.6 Retrieval and reranking procedure

At query time, the system parses the natural language description into a global query embedding and attribute mentions. It retrieves an initial candidate set by approximate nearest-neighbor search over image embeddings. A reranking stage then applies the composite attribute-aware score, enforces explicit negative constraints, and can incorporate spatial filters such as a study area or street-network buffer. The final output includes the ranked images, attribute-level evidence, and uncertainty scores. This structure keeps the expensive image encoding offline while allowing query-specific weighting and constraints at retrieval time. It also provides a natural point for human review, because the attribute decomposition shows why an image was ranked highly or rejected.

4. Analytical Assessment and Design Trade-offs

4.1 Dual encoder versus caption-then-retrieve

An alternative to dual encoding is to generate captions for every street view image and then apply text retrieval over the captions. Caption-then-retrieve can be interpretable and can leverage general-purpose language models, but it introduces error propagation: a caption that omits a small but decisive visual cue, such as a curb cut or a blank ground-floor facade, can make a relevant image undiscoverable. Captions also compress spatial relations into linear text, which is poorly suited to compositional queries about enclosure, frontage continuity, and street-level activity. Dual encoding avoids this bottleneck and supports efficient search, but it offers less explicit intermediate evidence. The attribute-aware formulation addresses this trade-off by adding interpretable similarity components without requiring a full caption for every image.

4.2 Ontology coverage and label noise

The ontology is both an asset and a source of constraint. A well-specified attribute set makes queries interpretable and supports sensitivity analysis, but it may fail to cover unanticipated descriptions. Scene-level datasets such as Places illustrate the breadth of visual scene categories [18], while semantic urban datasets such as Cityscapes show the value of pixel-level categories for roads, buildings, vegetation, and people [19]. Neither resource alone supplies the relational urban design qualities that natural language queries often target. The retrieval framework should therefore treat ontology labels as partial evidence. Soft attribute targets, multi-label annotations, and explicit unknown categories are preferable to forcing every image into a single scene class.

Label noise is particularly consequential for attributes that depend on context. A tree may indicate greenery, but a single street tree in a barren corridor does not produce the same canopy condition as a dense residential avenue. A bench may indicate social opportunity, but only when it is accessible and situated near activity. Training should distinguish attribute

presence from attribute intensity and from relational configuration. Without this distinction, the model may learn object detectors rather than scene-characteristic representations.

4.3 Viewpoint, spatial dependence, and audit validity

Street view images are sampled perspectives, not neutral measurements of a place. Camera height, direction, seasonal foliage, temporary construction, and privacy blurring affect what is visible. Spatial dependence further complicates evaluation: adjacent images often share buildings, trees, and street furniture, so random train-test splits can leak information and overstate generalization. Evaluation should therefore use spatial blocks or held-out cities, and it should report performance by scene type and attribute. Street view audits in public health research have long grappled with virtual observation validity [20]; retrieval systems inherit those concerns and add new ones related to ranking and query ambiguity.

4.4 Evaluation logic and error modes

Retrieval quality should be assessed with ranking metrics rather than classification accuracy alone. Mean average precision summarizes precision across the ranked list for a set of queries:

$$MAP = (1/|Q|) \sum_\{q \in Q\} (1/|R_q|) \sum_\{k = 1\}\^\{|R_q|\} P_q(k)$$

Here Q is the set of evaluation queries, R_q is the set of images judged relevant to query q, and P_q(k) is precision at rank k for that query. Recall at k, normalized discounted cumulative gain, and attribute-level precision provide complementary views. Error analysis should separate false negatives caused by incomplete relevance judgments from genuine ranking errors. It should also examine lexical shortcuts, such as a model ranking all images with visible trees highly regardless of the rest of the query; geographic shortcuts, such as memorizing neighborhood appearance; and attribute leakage, such as using sky openness as a proxy for enclosure without considering street width. These error modes are not incidental. They follow from the structure of urban imagery and from the many-to-many relation between language and scenes.

5. Technical Discussion and Implications

The framework has practical implications for urban analytics because it turns natural language into a searchable layer over street view archives. A planning analyst could retrieve examples of a described streetscape condition before designing a survey; a public health researcher could compare visible activity cues across neighborhoods; a community group could search for recurring visual concerns that are difficult to encode as a single indicator. The attribute decomposition supports deliberation, because users can see whether a result was driven by greenery, enclosure, frontage, or activity rather than by an uninterpretable global score. Uncertainty scores further allow the system to withhold automated judgments when visual evidence is weak.

The main boundary conditions are ontological and spatial. Retrieval can only be as expressive as the scene characteristics the model can ground in images, and natural language may refer to qualities that are not reliably visible from a single viewpoint. Spatial filtering can improve precision for local queries but may also reproduce existing inequalities in imagery coverage. Calibration requires relevance judgments that reflect the intended use, and those judgments may vary across users. These conditions do not invalidate the approach; they define the validation burden. A credible evaluation would need to report ranking performance,

calibration error, attribute-level sensitivity, and spatial generalization separately, rather than collapsing them into a single aggregate score.

6. Conclusion

This paper has developed a cross-modal retrieval framework for matching natural language descriptions of urban scene characteristics to street view images. The framework treats retrieval as a ranking problem, represents queries through a scene-characteristic ontology, aligns text and image encoders in a shared embedding space, and combines holistic similarity with attribute-specific evidence. Hard negative construction targets compositional and geographic shortcuts, while calibration separates relevance from confidence and supports abstention when visual evidence is ambiguous. The analytical assessment shows why dual encoding is preferable to caption-then-retrieve for compositional urban queries, why ontology labels should be treated as partial evidence, and why spatial blocking and attribute-level error analysis are necessary for credible evaluation. The contribution is not a claim of empirical superiority but a technically specified retrieval design that makes natural language a first-class modality for interrogating street view imagery. Such a design can support more transparent and flexible urban scene analysis while keeping the limits of visual inference visible to the user.

References

1. Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., & Fei-Fei, L. (2017). Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States. Proceedings of the National Academy of Sciences, 114(50), 13108–13113. https://doi.org/10.1073/pnas.1700035114

2. Biljecki, F., & Ito, K. (2021). Street view imagery in urban analytics and GIS: A review. Landscape and Urban Planning, 215, 104217. https://doi.org/10.1016/j.landurbplan.2021.104217

3. Kang, Y., Zhang, F., Gao, S., Lin, H., & Liu, Y. (2020). A review of urban physical environment sensing using street view imagery in public health studies. Annals of GIS, 26(3), 261–275. https://doi.org/10.1080/19475683.2020.1791954

4. Dubey, A., Naik, N., Parikh, D., Raskar, R., & Hidalgo, C. A. (2016). Deep learning the city: Quantifying urban perception at a global scale. In Computer Vision – ECCV 2016 (pp. 196–212). Springer. https://doi.org/10.1007/978-3-319-46448-0_12

5. Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H. H., Lin, H., & Ratti, C. (2018). Measuring human perceptions of a large-scale urban region using machine learning. Landscape and Urban Planning, 180, 148–160. https://doi.org/10.1016/j.landurbplan.2018.08.020

6. Lynch, K. (1960). The image of the city. MIT Press.

7. Ewing, R., & Handy, S. (2009). Measuring the unmeasurable: Urban design qualities related to walkability. Journal of Urban Design, 14(1), 65–84. https://doi.org/10.1080/13574800802451155

8. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (PMLR 139, pp. 8748–8763).

9. Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. T. (2020). Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 6769–6781). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.550

10. Aman, J., Matisziw, T. C., Kim, J. B., & Luo, D. (2022, April). Sensing the city: Leveraging geotagged social media posts and street view imagery to model urban streetscapes using deep neural networks. In Proceedings of the International Conference on Computer-Aided Architectural Design Research in Asia (pp. 595-604). Association for Computer-Aided Architectural Design Research in Asia.

11. Faghri, F., Fleet, D. J., Kiros, J. R., & Fidler, S. (2018). VSE++: Improving visual-semantic embeddings with hard negatives. In Proceedings of the British Machine Vision Conference (BMVC).

12. Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 815–823). https://doi.org/10.1109/CVPR.2015.7298682

13. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In Proceedings of the 37th International Conference on Machine Learning (PMLR 119, pp. 1597–1607).

14. Wang, L., Li, Y., Huang, J., & Lazebnik, S. (2019). Learning two-branch neural networks for image-text matching tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 394–407. https://doi.org/10.1109/TPAMI.2018.2797921

15. Jia, C., Yang, Y., Xia, Y., Chen, Y. T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y. H., Li, Z., & Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the 38th International Conference on Machine Learning (PMLR 139, pp. 4904–4916).

16. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (PMLR 70, pp. 1321–1330).

17. van den Oord, A., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv:1807.03748.

18. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(6), 1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009

19. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The Cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3213–3223). https://doi.org/10.1109/CVPR.2016.350

20. Rundle, A. G., Bader, M. D., Richards, C. A., Neckerman, K. M., & Teitler, J. O. (2011). Using Google Street View to audit neighborhood environments. American Journal of Preventive Medicine, 40(1), 94–100. https://doi.org/10.1016/j.amepre.2010.09.034