This paper reviews the use of machine learning in cancer prediction and prognosis, analyzing over 120 studies published between 1989 and 2006.

Key Findings:

* **Growing Trend:** The use of machine learning in cancer prediction and prognosis is rapidly increasing, with a 25% annual growth rate.
* **Data Types:** Most studies rely on clinical data (staging, histology, demographics), with an increasing trend toward incorporating molecular data (protein markers, gene mutations, gene expression).
* **Cancer Types:** Breast and prostate cancer dominate the studies, but machine learning is successfully applied to various other cancers.
* **Algorithms:** Artificial neural networks (ANNs) are the most commonly used algorithm, while support vector machines (SVMs) and decision trees are used less frequently.
* **Performance:** Machine learning methods generally improve prediction accuracy by 15-25% compared to conventional approaches.

Common Problems:

* **Lack of Validation:** Many studies lack sufficient internal or external validation, making it difficult to assess the reliability of the models.
* **Small Sample Size:** Some studies use small datasets, leading to overfitting and potentially unreliable results.
* **Sample-to-Feature Ratio:** Microarray studies often have too many features (genes) compared to samples, making the models susceptible to the "curse of dimensionality".
* **Data Quality:** Data entry errors and inconsistent feature selection can impact the accuracy of models.
* **Overreliance on ANNs:** The overreliance on ANNs, which are "black box" technologies, may hinder the adoption of more interpretable methods like decision trees and SVMs.

Recommendations:

* **Rigorous Validation:** Use n-fold cross-validation and external validation sets to ensure robustness and generalizability of models.
* **Adequate Sample Size:** Use sufficiently large datasets to avoid overfitting and ensure the sample-to-feature ratio is sufficient.
* **Data Quality Assurance:** Implement procedures for data entry verification and feature selection to ensure data quality.
* **Multiple Algorithms:** Explore multiple machine learning algorithms to find the optimal method for a particular problem.
* **Detailed Documentation:** Provide thorough methodological documentation, including detailed data descriptions and algorithm implementations, to enable replication and verification of results.

Overall: The review highlights the potential of machine learning to improve cancer prediction and prognosis, but emphasizes the importance of rigorous research methodology for developing reliable and transferable models.