

APPENDIX WITH SUPPORTING INFORMATION
– FOR ONLINE PUBLICATION ONLY –

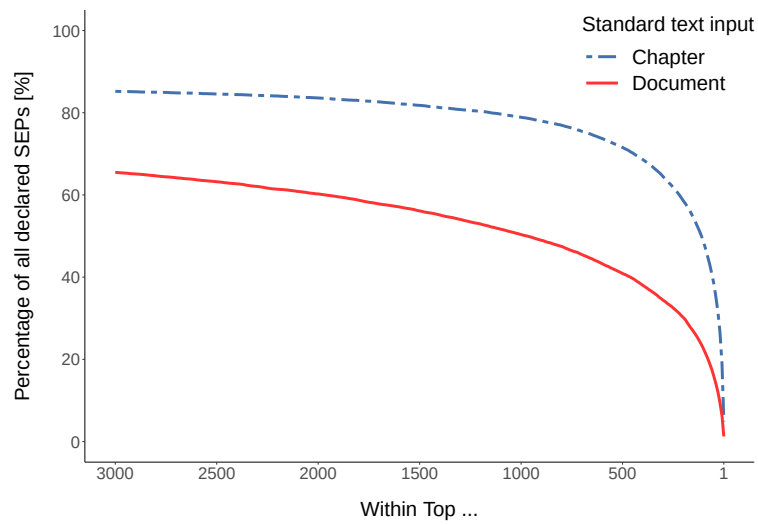
A Appendix: Figures

Figure A-1: Textual similarity between patents and standards (2nd example)

Patent publication: US 9,213,605 B2 (2015-12-15) <i>At operation 4020, upon receipt of a SIP INVITE request, the SCF examines the SDP parameters in the SDP offer and performs service authorization procedures to check the service rights of the requested MBMS download service according to the user subscription information.</i> <i>If the user is authorized, the SCF may answer with a SIP 200 OK including the SDP answer. The SDP answer may contain the fdt_address:uri to indicate the address of the File Delivery Table (FDT) and the repair-server-address:uri to indicate the address of the repair server.</i>	Standard specification: ETSI TS 126 237 V12.1.0 (2014-19) <i>"Upon receipt of SIP INVITE request, the SCF shall perform service authorization procedures to check the service rights of requested MBMS download service according to the user subscription information [...].</i> <i>If the SDP parameters are examined successfully, the SCF shall answer with a SIP 200 OK, indicating the SDP answer as follows:</i> <i>· The m-line(s) and c-line(s) shall be identical to ones indicated in the SDP offer.</i> <i>· An a=fdt_address:uri to indicate the address of the File Delivery Table.</i> <i>· An a=repair-server-address:uri to indicate the address of the repair server."</i>
---	--

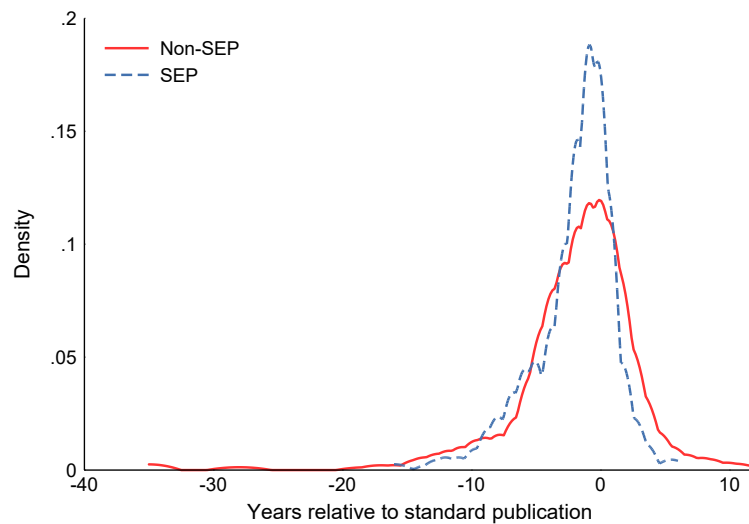
Notes: The quoted text of the patent publication is part of the patent description.

Figure A-2: Aggregate share of SEPs by rank



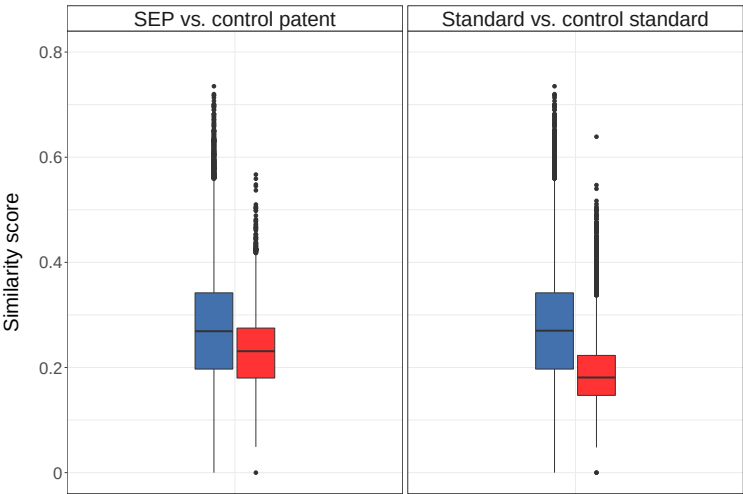
Notes: This graph shows the aggregate share of SEPs by their similarity rank at document level (red line) and chapter (blue line) level.

Figure A-3: Time difference between patent (priority) filing and standard publication for patents with high similarity



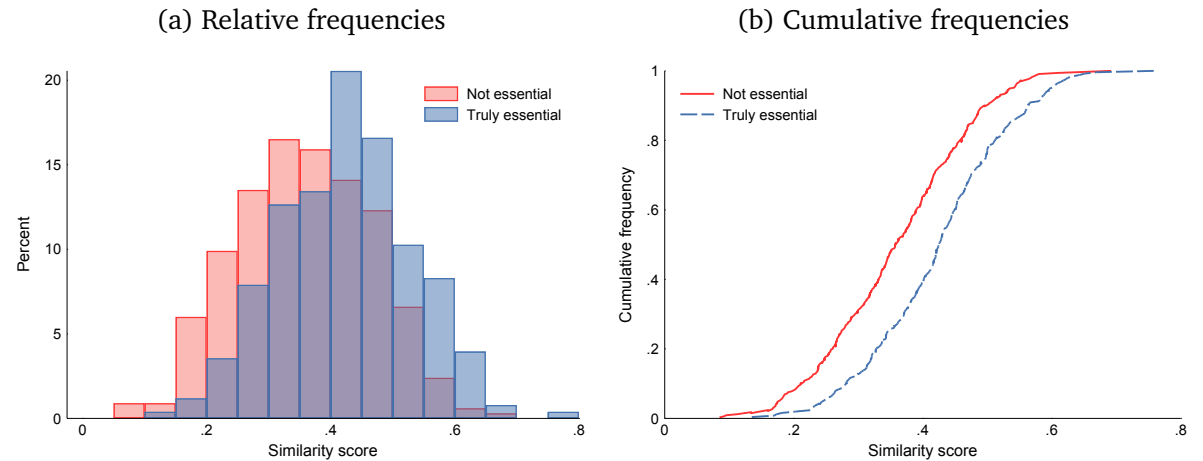
Notes: This graph shows the time difference between patent (priority) filing and standard publication in years relative to standard publication. We distinguish between SEPs (blue) and non-SEPs (red) with high similarity (i.e., a similarity score of at least 0.6 and a similarity rank of 5 or better).

Figure A-4: Comparison of SEP-standard pairs with control groups (censored data)



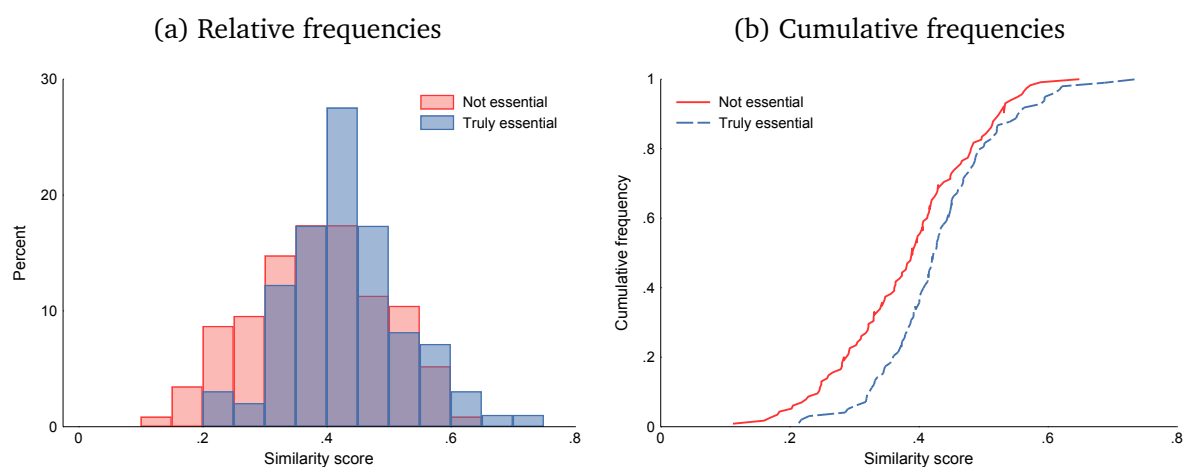
Notes: The box plot on the left-hand side shows the difference in similarity scores of SEP declarations (blue) and control patents compared to the same standard document (red). On the right-hand side, similarity scores of SEP declarations (blue) are compared to similarity scores of the same SEP and control standard documents (red). The censored dataset is used in this representation. Differences are significant but considerably less pronounced relative to the results when using the truncated data. Statistics are shown in Online Appendix Table B-2.

Figure A-5: Distribution of SEP similarity scores by essentiality assessment (UMTS standards)



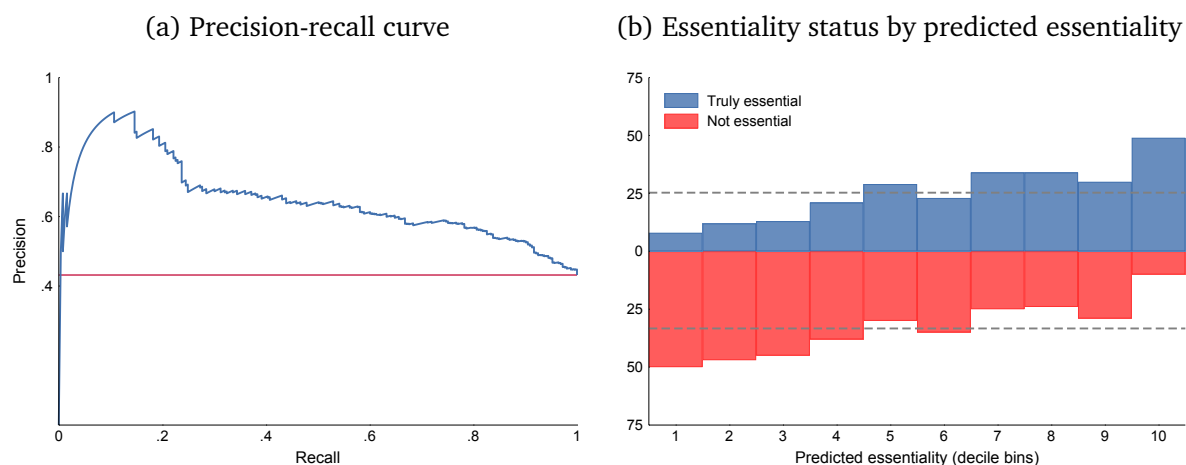
Notes: The right-hand graph shows the similarity score distributions of the two subsets of assessed SEPs declared to ETSI UMTS standards: not essential SEPs (red bars) and truly essential SEPs (blue bars). The left-hand graph shows cumulative frequencies for both subsets.

Figure A-6: Distribution of SEP similarity scores by essentiality assessment (GSM standards)



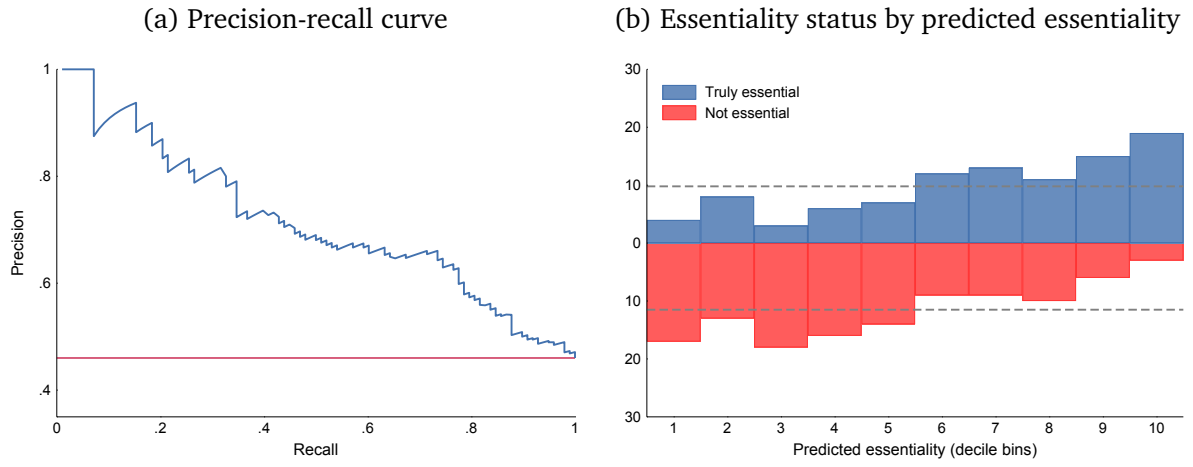
Notes: The right-hand graph shows the similarity score distributions of the two subsets of assessed SEPs declared to ETSI GSM standards: not essential SEPs (red bars) and truly essential SEPs (blue bars). The left-hand graph shows cumulative frequencies for both subsets.

Figure A-7: Predictive performance (UMTS standards)



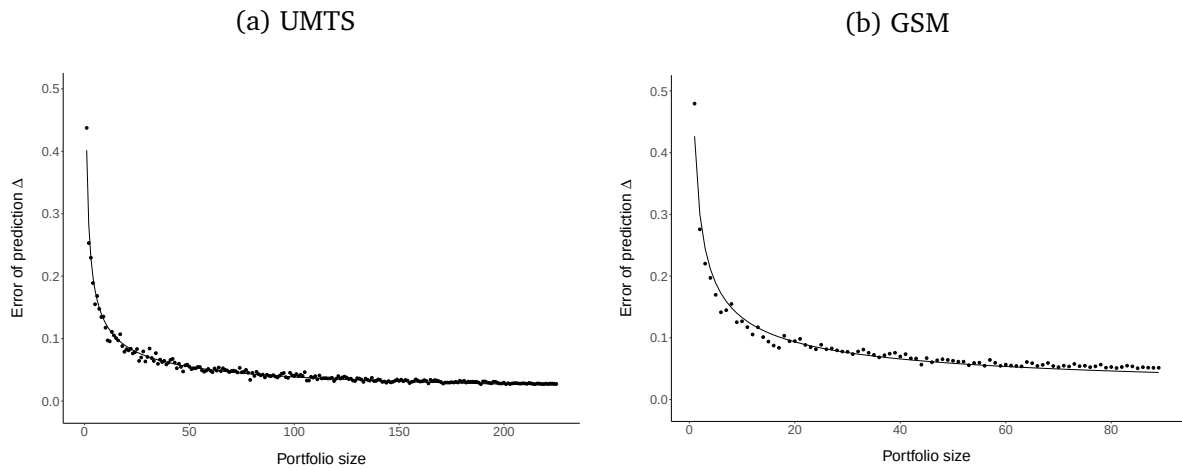
Notes: The left-hand graph illustrates the precision-recall curve. The right-hand graph visualizes the composition of SEPs by essentiality status within bins of the predicted probability of true essentiality. In both graphs, predictions are based on the specification in Online Appendix Table B-6, column 5.

Figure A-8: Predictive performance (GSM standards)



Notes: The left-hand graph illustrates the precision-recall curve. The right-hand graph visualizes the composition of SEPs by essentiality status within bins of the predicted probability of true essentiality. In both graphs, predictions are based on the specification in Online Appendix Table B-7, column 5.

Figure A-9: The error of prediction as a function of portfolio size



Notes: The error of prediction Δ is plotted as a function of portfolio size where portfolios are randomly drawn from the test sample of UMTS and GSM patents. Non-linear least squares fits are shown. The fitted functions are power law functions.

B Appendix: Tables

Table B-1: Patent characteristics of SEPs and non-SEPs with high similarity to standards

High similarity patents	SEPs (N = 186)			Non-SEPs (N = 239)			Diff.	p-value
	Mean	Median	Std. Dev.	Mean	Median	Std. Dev.		
Priority year	2006.70	2007.00	3.73	2006.26	2008.00	5.44	-0.44	0.342
Patent family size	9.62	7.00	7.88	5.04	4.00	3.47	-4.59	0.000***
# Applicants	1.83	1.00	1.39	1.78	1.00	1.44	-0.05	0.717
# Inventors	2.46	2.00	1.35	2.46	2.00	1.35	0.00	0.996
Corporate applicant	0.94	1.00	0.25	0.86	1.00	0.35	-0.07	0.014**
SEP-holding applicant	1.00	1.00	0.00	0.84	1.00	0.37	-0.16	0.000***
# Patent references	19.23	13.00	28.45	15.03	11.00	21.81	-4.20	0.086*
# NPL references	23.73	10.00	44.40	8.47	4.00	20.32	-15.25	0.000***
# Claims	18.87	18.00	10.97	20.58	20.00	11.18	1.71	0.138
Length claim 1	110.53	98.00	62.00	116.06	97.00	62.79	5.53	0.394
# US fwd. cit. (5yrs)	34.38	18.00	47.34	25.08	14.00	30.78	-9.31	0.017**
# SEP US fwd. cit. (5yrs)	6.95	3.00	10.20	3.24	1.00	6.07	-3.71	0.000***
Patent transferred	0.04	0.00	0.20	0.02	0.00	0.13	-0.03	0.105
Years betw priority filing and std publ.	2.56	2.00	2.80	3.48	2.00	4.61	0.91	0.018**

Notes: This table compares patent characteristics between SEPs and non-SEPs with high similarity (i.e., a similarity score of at least 0.6 and a similarity rank of 5 or better. The unit of observation is at the patent level. Reported p -values based on an unpaired t -test. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table B-2: T-statistics for the comparison of SEP-standard pairs with control groups

	t-value	Δ Score
Uncensored		
SEP vs. control patent	100***	0.097
Standard vs. control standard	151***	0.147
Censored		
SEP vs. control patent	68***	0.049
Standard vs. control standard	126***	0.090

Notes: Δ Score denotes the differences in mean similarity scores for both groups. Significance levels: *** $p < 0.01$.

Table B-3: Summary statistics (assessed SEP regression sample for LTE standards)

	Mean	SD	Median	Min	Max	N
Dependent variable:						
Truly essential to LTE	0.381	0.486	0	0	1	1241
Independent variables:						
Similarity score	0.359	0.106	0	0	1	1241
Similarity score (embeddings)	0.567	0.055	1	0	1	1241
Similarity score (tf-idf)	0.250	0.134	0	0	1	1241
- no desc	0.194	0.118	0	0	1	1241
- claims	0.188	0.117	0	0	1	1241
- early pub	0.245	0.132	0	0	1	1241
- no desc early pub	0.182	0.109	0	0	1	1241
- claims early pub	0.176	0.108	0	0	1	1241
Priority year	2006.114	3.593	2007	1991	2012	1241
Patent family size	11.148	9.278	9	1	120	1241
# Inventors	3.179	1.781	3	1	13	1241
# Applicants	2.258	1.911	1	1	13	1241
# Independent claims	4.372	2.859	4	1	20	1241
Length claim 1	107.546	58.398	96	1	449	1241
# Patent references	24.944	36.970	17	1	911	1241
# NPL references	26.943	62.166	11	0	1188	1241
# SEP US fwd. cit. (5yrs)	8.623	11.501	5	0	138	1241
SEP transferred	0.065	0.247	0	0	1	1241
Section-specific decl.	0.369	0.483	0	0	1	1241
Earliest decl. year	2010.699	2.641	2011	2001	2016	1241

Notes: Summary statistics for the regression sample of assessed SEPs declared to LTE standards.

Table B-4: Summary statistics (assessed SEP regression sample for UMTS standards)

	Mean	SD	Median	Min	Max	N
Dependent variable:						
Truly essential to UMTS	0.432	0.496	0	0	1	586
Independent variables:						
Similarity score	0.387	0.113	0	0	1	586
Similarity score (embeddings)	0.704	0.053	1	0	1	586
Similarity score (tf-idf)	0.293	0.145	0	0	1	586
- no desc	0.214	0.122	0	0	1	586
- claims	0.205	0.121	0	0	1	586
- early pub	0.289	0.145	0	0	1	586
- no desc early pub	0.203	0.119	0	0	1	586
- claims early pub	0.195	0.117	0	0	1	586
Priority year	2003.556	3.802	2004	1995	2012	586
Patent family size	13.802	11.735	11	1	156	586
# Inventors	2.700	1.608	2	1	10	586
# Applicants	1.881	1.628	1	1	12	586
# Independent claims	4.498	3.564	4	1	33	586
Length claim 1	117.935	70.744	102	1	489	586
# Patent references	27.068	29.913	19	0	314	586
# NPL references	27.222	72.429	9	0	1188	586
# SEP US fwd. cit. (5yrs)	6.923	9.378	4	0	103	586
SEP transferred	0.087	0.282	0	0	1	586
Section-specific decl.	0.396	0.489	0	0	1	586
Earliest decl. year	2008.345	3.482	2009	2001	2014	586

Notes: Summary statistics for the regression sample of assessed SEPs declared to UMTS standards.

Table B-5: Summary statistics (assessed SEP regression sample for GSM standards)

	Mean	SD	Median	Min	Max	N
Dependent variable:						
Truly essential to GSM	0.460	0.500	0	0	1	213
Independent variables:						
Similarity score	0.407	0.104	0	0	1	213
Similarity score (embeddings)	0.718	0.050	1	1	1	182
Similarity score (tf-idf)	0.305	0.144	0	0	1	182
- no desc	0.202	0.116	0	0	1	182
- claims	0.194	0.114	0	0	1	182
- early pub	0.302	0.147	0	0	1	182
- no desc early pub	0.197	0.116	0	0	1	182
- claims early pub	0.189	0.115	0	0	1	182
Priority year	2003.746	4.049	2004	1996	2011	213
Patent family size	12.850	9.981	11	1	94	213
# Inventors	2.671	1.562	2	1	11	213
# Applicants	2.127	1.829	1	1	12	213
# Independent claims	4.728	3.265	4	1	18	213
Length claim 1	119.009	62.000	109	1	373	213
# Patent references	25.714	28.297	17	0	197	213
# NPL references	29.047	65.825	9	0	511	213
# SEP US fwd. cit. (5yrs)	4.671	5.625	3	0	29	213
SEP transferred	0.113	0.317	0	0	1	213
Section-specific decl.	0.488	0.501	0	0	1	213
Earliest decl. year	2008.540	3.521	2009	2001	2014	213

Notes: Summary statistics for the regression sample of assessed SEPs declared to GSM standards.

Table B-6: Logistic regressions: UMTS standard essentiality

	DV: SEP truly essential (d)				
	(1)	(2)	(3)	(4)	(5)
Similarity score	1.3517*** (0.1954)	1.4141*** (0.2085)	1.8189*** (0.2767)	1.6212*** (0.3362)	1.3821*** (0.2045)
Patent characteristics	No	Yes	Yes	Yes	Lasso
Priority year FE	No	No	Yes	Yes	Lasso
Earliest decl. year FE	No	No	Yes	Yes	Lasso
CPC-4 FE	No	No	Yes	Yes	Lasso
Patent holder FE	No	No	No	Yes	No
Pseudo R^2	0.063	0.085	0.183	0.276	0.114
AUC	0.663	0.686	0.771	0.833	0.724
Observations	586	586	586	562	586

Notes: The dependent variable is binary and equal to one if the SEP is truly essential for UMTS standards as judged by SEP assessment. Marginal effects of one unit change are reported. AUC = Area under ROC curve. The sample size is fixed in all specifications to ease comparison of coefficients from different models. Standard errors in parentheses. Information on the control variables can be found in Online Appendix Table B-4. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table B-7: Logistic regressions: GSM standard essentiality

	DV: SEP truly essential (d)				
	(1)	(2)	(3)	(4)	(5)
Similarity score	1.1891*** (0.3515)	1.3266*** (0.4051)	0.7918* (0.4257)	1.0666 (0.8779)	1.1971*** (0.3660)
Patent characteristics	No	Yes	Yes	Yes	Lasso
Priority year FE	No	No	Yes	Yes	Lasso
Earliest decl. year FE	No	No	Yes	Yes	Lasso
CPC-4 FE	No	No	Yes	Yes	Lasso
Patent holder FE	No	No	No	Yes	No
Pseudo R^2	0.041	0.101	0.183	0.379	0.127
AUC	0.624	0.706	0.800	0.873	0.731
Observations	213	213	213	179	213

Notes: The dependent variable is binary and equal to one if the SEP is truly essential for GSM standards as judged by SEP assessment. Marginal effects of one unit change are reported. AUC = Area under ROC curve. The sample size is fixed in all specifications to ease comparison of coefficients from different models. Standard errors in parentheses. Information on the control variables can be found in Online Appendix Table B-5. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table B-8: Logistic regressions: UMTS standard essentiality (alternative algorithms)

	DV: SEP truly essential (d)				
	(1)	(2)	(3)	(4)	(5)
Similarity score (tf-idf)	0.9753*** (0.1495)	1.0112*** (0.1576)	1.1064*** (0.1850)	0.9773*** (0.2086)	0.9269*** (0.1522)
Patent characteristics	No	Yes	Yes	Yes	Lasso
Priority year	No	No	Yes	Yes	Lasso
Earliest decl. year FE	No	No	Yes	Yes	Lasso
CPC-4 FE	No	No	Yes	Yes	Lasso
Patent holder FE	No	No	No	Yes	No
Pseudo R^2	0.056	0.078	0.172	0.274	0.088
AUC	0.664	0.692	0.772	0.831	0.707
Observations	586	586	586	562	586
	(6)	(7)	(8)	(9)	(10)
Similarity score (embeddings)	1.0370*** (0.3937)	0.9729** (0.4015)	0.9850** (0.4668)	0.9526* (0.5396)	0.9683** (0.4147)
Patent characteristics	No	Yes	Yes	Yes	Lasso
Priority year	No	No	Yes	Yes	Lasso
Earliest decl. year FE	No	No	Yes	Yes	Lasso
CPC-4 FE	No	No	Yes	Yes	Lasso
Patent holder FE	No	No	No	Yes	No
Pseudo R^2	0.009	0.032	0.130	0.248	0.104
AUC	0.562	0.619	0.732	0.818	0.709
Observations	586	586	586	562	586

Notes: The dependent variable is binary and equal to one if the SEP is truly essential for UMTS standards as judged by SEP assessment. *Similarity score (tf-idf)* and *Similarity score (embeddings)* are based on two alternative open source algorithms: tf-idf and embeddings. For more information on the algorithms, see Section 3.2. Marginal effects of one unit change are reported. AUC = Area under ROC curve. The sample size is fixed in all specifications to ease comparison of coefficients from different models. Standard errors in parentheses. Information on the control variables can be found in Online Appendix Table B-4. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table B-9: Logistic regressions: GSM standard essentiality (alternative algorithms)

	DV: SEP truly essential (d)				
	(1)	(2)	(3)	(4)	(5)
Similarity score (tf-idf)	0.8287*** (0.2662)	0.8792*** (0.2813)	1.2435*** (0.4478)	7.4140 (78.4059)	0.8377*** (0.2753)
Patent characteristics	No	Yes	Yes	Yes	Lasso
Priority Year FE	No	No	Yes	Yes	Lasso
Earliest Decl. year FE	No	No	Yes	Yes	Lasso
CPC-4 FE	No	No	Yes	Yes	Lasso
Firm FE	No	No	No	Yes	No
Pseudo R^2	0.040	0.100	0.340	0.629	0.105
AUC	0.641	0.705	0.860	0.954	0.702
Observations	182	182	178	141	182
	(6)	(7)	(8)	(9)	(10)
Similarity score (embeddings)	1.7992** (0.7553)	1.7363** (0.7857)	1.4663 (1.2329)	0.2748 (2.7226)	1.6932** (0.7822)
Patent characteristics	No	Yes	Yes	Yes	Lasso
Priority Year FE	No	No	Yes	Yes	Lasso
Earliest Decl. year FE	No	No	Yes	Yes	Lasso
CPC-4 FE	No	No	Yes	Yes	Lasso
Firm FE	No	No	No	Yes	No
Pseudo R^2	0.023	0.079	0.311	0.543	0.086
AUC	0.597	0.697	0.844	0.931	0.672
Observations	182	182	178	141	182

Notes: The dependent variable is binary and equal to one if the SEP is truly essential for GSM standards as judged by SEP assessment. *Similarity score (tf-idf)* and *Similarity score (embeddings)* are based on two alternative open source algorithms: tf-idf and embeddings. For more information on the algorithms, see Section 3.2. Marginal effects of one unit change are reported. AUC = Area under ROC curve. The sample size is fixed in all specifications to ease comparison of coefficients from different models. Standard errors in parentheses. Information on the control variables can be found in Online Appendix Table B-5. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table B-10: Logistic regressions: LTE standard essentiality (patent publication and patent text)

	DV: SEP truly essential (d)					
	(1)	(2)	(3)	(4)	(5)	(6)
	Latest			Earliest		
Patent publication:						
Patent text:	Full	No desc	Claims	Full	No desc	Claims
Similarity score (tf-idf)	0.6266*** (0.1158)	0.4832*** (0.1328)	0.4328*** (0.1316)	0.5719*** (0.1195)	0.3706*** (0.1438)	0.3710** (0.1452)
Controls	Lasso	Lasso	Lasso	Lasso	Lasso	Lasso
Pseudo R^2	0.120	0.122	0.118	0.120	0.112	0.112
AUC	0.727	0.729	0.726	0.726	0.720	0.720
Observations	1,241	1,241	1,241	1,241	1,241	1,241

Notes: The dependent variable is binary and equal to one if the SEP is truly essential for LTE standards as judged by SEP assessment. The similarity (*Similarity score (tf-idf)*) is based on different text input: different patent publications and input text. Concerning the patent publication, either the latest publication (*Latest*) or the earliest publication (*Earliest*) in the DOCDB patent family is chosen. Concerning the patent text either full text (*Full*), full text without description (*No desc*), or claims text only (*Claims*) is chosen. Marginal effects of one unit change are reported. AUC = Area under ROC curve. The sample size is fixed in all specifications to ease comparison of coefficients from different models. Standard errors in parentheses. Control variable categories are collapsed to one (*Controls*), but the specification is equivalent to the one in Online Appendix Table 4 Column 5. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table B-11: Logistic regressions: UMTS standard essentiality (patent publication and patent text)

	DV: SEP truly essential (d)					
	(1)	(2)	(3)	(4)	(5)	(6)
	Latest			Earliest		
Patent publication:						
Patent text:	Full	No desc	Claims	Full	No desc	Claims
Similarity score (tf-idf)	0.9269*** (0.1522)	0.9764*** (0.1936)	0.9974*** (0.1964)	0.9081*** (0.1501)	0.8723*** (0.1971)	0.8767*** (0.2008)
Controls	Lasso	Lasso	Lasso	Lasso	Lasso	Lasso
Pseudo R^2	0.088	0.134	0.140	0.084	0.135	0.135
AUC	0.707	0.740	0.746	0.704	0.743	0.742
Observations	586	586	586	586	586	586

Notes: The dependent variable is binary and equal to one if the SEP is truly essential for UMTS standards as judged by SEP assessment. The similarity (*Similarity score (tf-idf)*) is based on different text input: different patent publications and input text. Concerning the patent publication, either the latest publication (*Latest*) or the earliest publication (*Earliest*) in the DOCDB patent family is chosen. Concerning the patent text either full text (*Full*), full text without description (*No desc*), or claims text only (*Claims*) is chosen. Marginal effects of one unit change are reported. AUC = Area under ROC curve. The sample size is fixed in all specifications to ease comparison of coefficients from different models. Standard errors in parentheses. Control variable categories are collapsed to one (*Controls*), but the specification is equivalent to the one in Online Appendix Table B-8 Column 5. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table B-12: Logistic regressions: GSM standard essentiality (patent publication and patent text)

	DV: SEP truly essential (d)					
	(1)	(2)	(3)	(4)	(5)	(6)
	Latest			Earliest		
Patent publication:						
Patent text:	Full	No desc	Claims	Full	No desc	Claims
Similarity score (tf-idf)	0.8377*** (0.2753)	0.8313** (0.3406)	0.8189** (0.3443)	0.9304*** (0.2745)	0.5082 (0.3200)	0.4995 (0.3236)
Controls	Lasso	Lasso	Lasso	Lasso	Lasso	Lasso
Pseudo R^2	0.105	0.097	0.096	0.115	0.043	0.042
AUC	0.702	0.691	0.688	0.711	0.627	0.624
Observations	182	182	182	182	182	182

Notes: The dependent variable is binary and equal to one if the SEP is truly essential for GSM standards as judged by SEP assessment. The similarity (*Similarity score (tf-idf)*) is based on different text input: different patent publications and input text. Concerning the patent publication, either the latest publication (*Latest*) or the earliest publication (*Earliest*) in the DOCDB patent family is chosen. Concerning the patent text either full text (*Full*), full text without description (*No desc*), or claims text only (*Claims*) is chosen. Marginal effects of one unit change are reported. AUC = Area under ROC curve. The sample size is fixed in all specifications to ease comparison of coefficients from different models. Standard errors in parentheses. Control variable categories are collapsed to one (*Controls*), but the specification is equivalent to the one in Online Appendix Table B-9 Column 5. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table B-13: Confusion matrix (LTE standards)

Prediction	True SEPs	
	No	Yes
	No	Yes
No	233	104
Yes	46	53

Notes: Confusion matrix for the test set of LTE SEPs evaluated by the manual SEP assessments data. Results based on the lasso model specification (Table 3, column 5).

Table B-14: Predictive performance overview

Model	Similarity score			Similarity score (tf-idf)		
	Precision	Recall	F1	Precision	Recall	F1
Logit	0.65	0.66	0.65	0.64	0.66	0.64
NB	0.62	0.63	0.57	0.60	0.63	0.57
SVM (Linear)	0.64	0.65	0.59	0.63	0.64	0.57
SVM (Poly)	0.64	0.65	0.60	0.63	0.65	0.59
SVM (Radial)	0.63	0.64	0.60	0.64	0.65	0.61
XGB	0.64	0.65	0.63	0.64	0.65	0.63
RF	0.63	0.64	0.59	0.63	0.64	0.60

Notes: We use the Lasso specification (Table 3, column 5) to compare five different machine learning models for the prediction of true standard essentiality. The logistic classifier *Logit* is our main model and yields the highest scores for all three performance metrics. *NB* is the Naive Bayes classifier and *SVM* is the support vector machine. To account for potential non-linear separability, we use *radial* and *polynomial* kernel functions for the *SVM* classifier. We further report the performance of two ensemble algorithms: *RF* is a random forest and *XGB* is the Extreme Gradient Boosting algorithm. Scores are weighted according to the frequency of the corresponding classes.